

S

Sand Skink

- ▶ Friction-Reducing Sandfish Skin

SANS

- ▶ Small-Angle Scattering of Nanostructures and Nanomaterials

SAXS

- ▶ Small-Angle Scattering of Nanostructures and Nanomaterials

Scanning Electron Microscopy

Yimei Zhu¹ and Hiromi Inada²

¹Center for Functional Nanomaterials,
Brookhaven National Lab, Upton, NY, USA
²Hitachi High-Technologies Corporation,
Hitachinaka, Japan, Japan

Synonyms

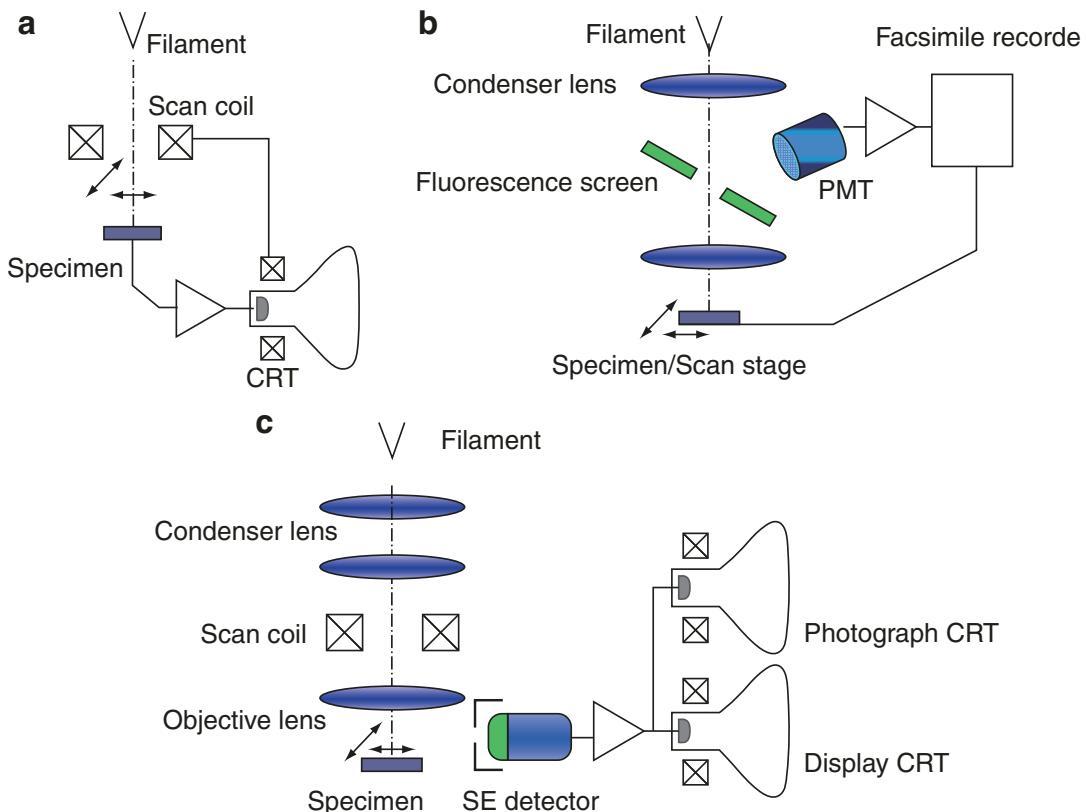
Scanning electron microscopy and secondary-electron imaging microscopy

Definition

A scanning electron microscope (SEM) is an instrument that uses high-energy electrons in a raster scan pattern to form images, or collect other signals, from the three-dimensional surface of a sample.

Introduction

The scanning electron microscope (SEM) is one of the most popular and user-friendly imaging tools that reveal the surface topography of a sample. It is also widely used for structural characterization of materials and devices, especially in the field of nanotechnology. Today there are in excess of 50,000 SEMs worldwide, and it is often seen as a “must-have” apparatus for research institutes and industry laboratories. In the SEM, incident electrons interact with the atoms that make up the sample producing signals that contain information about the sample’s surface morphology, composition, and other physical and chemical properties. The most common imaging mode in SEM lies in using secondary electrons. Since secondary electrons have very low energies, they are generated in and escape from regions near the sample surface. Combined with various detection systems, a SEM also can be used to determine the sample’s chemical composition through energy-dispersive x-ray spectroscopy and Auger electron spectroscopy and identify its phases through



Scanning Electron Microscopy, Fig. 1 Schematics of configurations of the scanning electron microscopes proposed and developed by some notable pioneers in the SEM

history. (a) Knoll in 1935, (b) Zworykin in 1945, and (c) Cambridge Instruments in 1965

analyzing electron diffraction patterns, mostly via high-energy backscattered electrons. Besides backscattered electrons, Auger electrons, characteristic x-rays, and other signals are generated from the interactions of the incident beam, and the sample under a SEM includes plasmons, bremsstrahlung radiation (noncharacteristic x-rays), cathodoluminescence, and electron-beam-induced current. This entry is focused on secondary-electron imaging related instrumentation, signal generation processes, and state-of-the-art imaging capabilities in SEM.

A Brief History

SEM was invented by Max Knoll in 1935 in Germany [1, 2] to study the targets of television tubes. The instrument consisted of electron-beam

deflection coils that scan the beam on a plate as the sample in a cathode ray tube (CRT) and an amplifier that boosts the plate current to display the signal on another CRT (Fig. 1a). Two years later, Manfred von Ardenne built an electron microscope with a highly demagnified probe using two condenser lenses for scanning transmission electron microscopy and also tried it as a SEM. Zworykin and his coworkers of the RCA Laboratories in the USA designed and built a dedicated SEM in 1942. Its electron optics includes three electrostatic lenses with scan coils placed between the second and third ones. A photomultiplier was first used to detect secondary electrons (Fig. 1b). The essential components of this apparatus are similar to those used in modern SEMs. The probe size of the incident, or primary, electron beam had a diameter of about 10 nm. However, compared with transmission

electron microscopes (TEMs), at that period, it could not image secondary electrons satisfactorily due to the poor signal-to-noise ratios of the images [3]. Sir Charles Oatley and Dennis McMullan built their first SEM at Cambridge University in 1948. The SEM technology was further pioneered by many postgraduate students at Cambridge including Gary Stewart. The first commercial instrument, named as “Stereoscan,” was launched in 1965 by the Cambridge Scientific Instrument Company for DuPont [4]. The instrument consists of electron multiplier detector with beryllium-copper dynodes to detect scattered electrons from the specimen surface. Images were displayed on a CRT, while another synchronized CRT recorded them on camera film. An Everhart-Thornley-type secondary-electron detector [5] significantly improved the detection efficiency of the low-energy secondary-electron signals (Fig. 1c). JEOL produced first commercial Japanese SEM, JSM-1, in 1966, while Hitachi commercialized its SEM, HSM-2, in 1969.

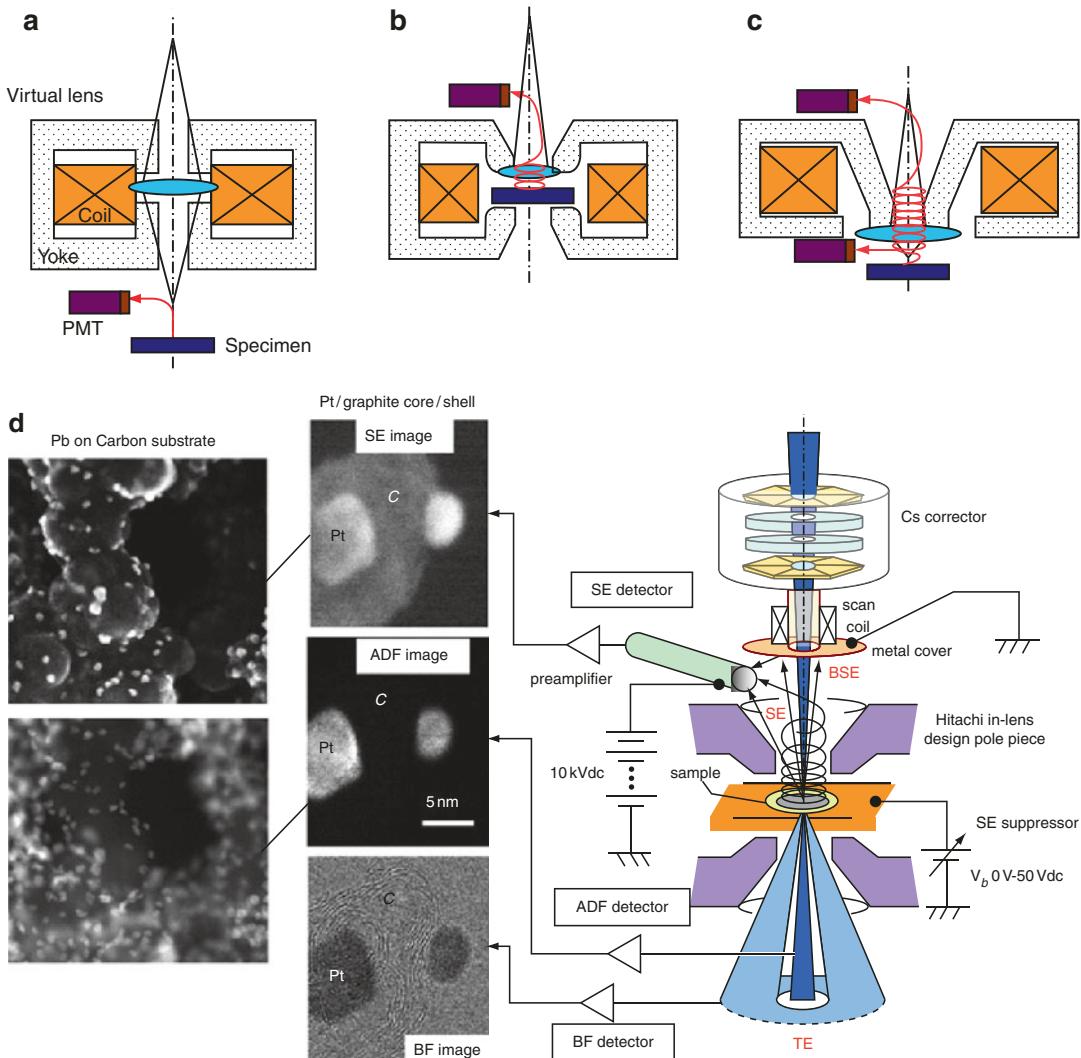
Instrumentation

A typical SEM consists of an electron gun, an electron lens system, various electron-beam deflection coils, electron detectors, and display and recording devices [6]. In a SEM, an electron beam is emitted from an electron gun sitting on a thermionic filament cathode, or from a field-emission needle tip. Tungsten is often used for thermionic electron guns due to its low cost, high melting point, and low vapor pressure so that it tolerates heating to about 2,800 K for electron emission. Other types of electron emitters include lanthanum hexaboride (LaB_6) cathodes, which offer higher brightness but require a better vacuum to avoid oxidizing the gun. Field-emission guns (FEGs) often used in modern SEMs can be the thermally assisted Schottky type, using emitters of zirconium oxide (ZrO), or the cold cathode type using tungsten <310> single crystal emitters and operated at room temperature. A cold field-emission gun has a much smaller source size (5–10 nm) than a

tungsten filament (1–10 μm) with three to four orders of magnitude larger current density and brightness [7].

The typical energy range of the electron beam used in SEM is from 0.5 to 40 keV. The electron-condenser lens system usually demagnifies the electron source more than hundreds of times to form a small probe on the sample. The beam passes through pairs of deflection coils, or scanning coils, in the electron column, typically in the final lens, which deflect the beam in the x and y axes by applying an incremental current into the scan coils, so that it scans in a raster fashion over a rectangular area of the sample surface.

Historically, SEM incorporates an objective lens. Unlike the objective lens in optical microscopes or transmission electron microscopes (TEMs), its purpose in SEM is not to image the sample but to focus the small probe on the sample. There are three types of objective lenses: out-lens (Fig. 2a), in-lens (Fig. 2b), and semi-in-lens (Fig. 2c). Most early SEMs had the simplest out-lens design, in which the sample sits beneath the lens leaving a large area available in the sample chamber. However, the yoke gap across the optical axis acts as a lens with leakage field, thus yielding significant imaging aberration. The in-lens pole piece, originally designed for TEMs, was adopted for SEM to reduce spherical aberration. Hitachi developed the first commercial SEM with such a design in 1985 [8]. Their microscope operating at 30 kV reaches a probe size of 0.5 nm with spherical aberration coefficient C_s of 1.6 mm. The drawback of the design is that a conventional thin sample, similar to that used in TEM, is required because the sample sits inside the pole-piece gap. A design resulting from a compromise between the out-lens and in-lens is the semi-in-lens pole piece (Fig. 2c), offering reasonable spatial resolution but with larger open space for the sample chamber so that a thick sample can be used. Such a design allows the sample to be placed a few mm from the pole piece and a large magnetic field to be applied to produce a smaller focal length and less spherical aberration. SEMs with the semi-in-lens design played a significant role in characterizing devices for the semiconductor industry in early 1990s.



Scanning Electron Microscopy, Fig. 2 (a–c) Three different objective lens designs. (a) Out-lens, (b) in-lens, and (c) semi-in-lens. (d) Schematic of the lens detector configuration of the aberration-corrected scanning electron microscope, Hitachi HD2700C, that routinely achieves an atomic resolution in imaging using secondary electrons and transmitted electrons. Two examples are shown on

the left; one is Pb particles on a carbon support, and the other is Pt particles with carbon-graphite shells. BF bright-field, ADF annular dark-field, SE secondary electrons, TE transmitted electrons, and BES backscattered electrons. The SE images clearly give a topographic view of the area and higher brightness of the light element C, compared with the corresponding ADF images

It is noteworthy that the in-lens design shown in Fig. 2b is very similar to that used in conventional TEM and/or STEM (scanning transmission electron microscope). Thus, with an efficient secondary-electron detector, a TEM or a STEM also can image a sample surface [9, 10]. Figure 2d shows a schematic of the objective lens (in-lens

design), the sample, and the arrangement of detectors in the Hitachi HD2700C aberration-corrected SEM/STEM [11]. The instrument operates at 80–200 kV, allowing simultaneous acquisition of bright-field (BF), annular dark-field (ADF), and secondary-electron (SE) images. In the BF and ADF modes, the transmitted electrons are used

to form the images, thus providing structural information from the sample's interior. In contrast, in the SE mode, electrons emerging from the surface with low energies, or short escape length, are used, and thus, the signals are surface sensitive. Different imaging modes have their own advantages and limitations. Since the BF signals are close to the phase contrast seen in TEM, they offer high spatial resolution, but are difficult to interpret. ADF images, on the other hand, are based on Rutherford scattering, and thus, their image intensity is directly related to the sample's atomic number Z (the so-called Z-contrast imaging). Since BF and ADF images are projected images, they give little structural information in the direction of the beam's trajectory. In contrast, SE imaging offers depth information on the surface topography and is more sensitive to the light elements than is ADF imaging (see Fig. 2d). Combining these different imaging modes in an electron microscope with an in-lens design, it has been demonstrated that simultaneously imaging both surface (SE) and bulk (ADF) at atomic resolution is possible in thin samples for a wide range of elements, from uranium and gold to silicon and carbon [11, 12].

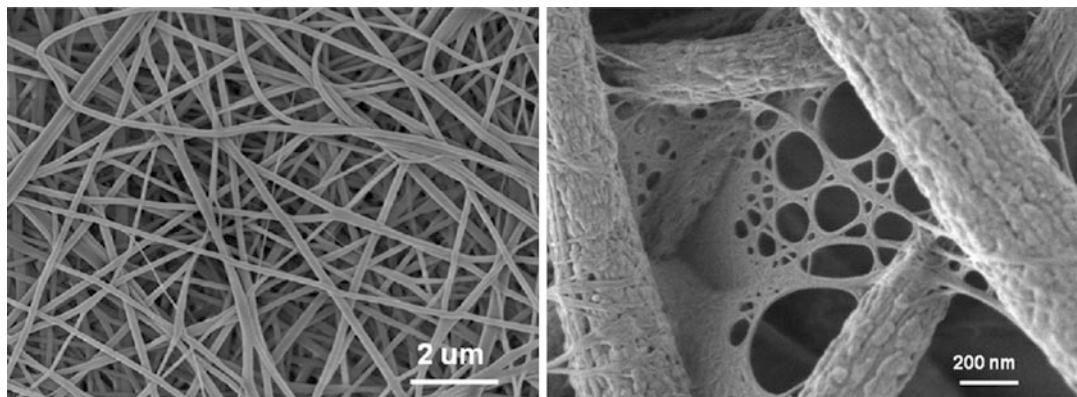
The electron detector is another important part of the SEM instrumentation. Although detector itself does not determine the image resolution in SEMs, it is essential to improve the SEM resolving power in terms of the signal-to-noise ratio. The commonest detector used in SEMs today still is that developed by Everhart and Thornley in 1957 [5], the so-called E-T detector. The detector consists of a photomultiplier, a light guide, and a positively biased scintillator. To attract low-energy electrons effectively, the scintillator is applied in a 10 kV dc bias to accelerate the electrons. The energized electrons cause the scintillator to emit flashes of light (cathodoluminescence) that then are transmitted to the photomultiplier. The amplified output of the electrical signals by the photomultiplier is displayed as a two-dimensional intensity distribution that can be viewed and photographed on an analogue video display. The Everhart-Thornley detector, which normally is positioned to one side above the specimen, exhibits low efficiency in detecting backscattered electrons

because few such electrons are emitted in the solid angle subtended by the detector. Furthermore, its positive bias cannot readily attract the high-energy backscattered electrons (close to the energy of the incident electrons). Backscattered electrons are usually collected above the sample in a "doughnut-type" arrangement, concentric with the electron beam, to maximize the solid angle of collection.

Secondary-Electron Signal Generation

Secondary-electron (SE) imaging is the most frequently used mode of imaging in SEM. Secondary electrons, defined as the electrons with energy below 50 eV, are generated along the primary electrons' trajectories within the sample, but are subject to elastic and inelastic scattering during their passage through the sample. These electrons can be valence electrons or are ejected from the orbits of the inner shells (most likely the k-shells) of the sample. The consequence of their low kinetic energy is their shallow escape depth, which is about 1 nm for metals and up to 10 nm for insulators. The probability of escape decreases continuously with the increase of the depth below the surface. However, there is a nonzero probability of secondary-electron emission arising from inelastic scattering below the escape depth, as, for example, when a primary electron creates a fast secondary electron (energy > 50 eV) that travels toward the surface and generates a lower-energy secondary electron within the escape depth. The production of the secondary electron signals involves the generation, propagation, escape from the surface, and arrival at the detector. These four processes are detailed in the atomic imaging section.

The secondary electrons discussed here are often referred to as SE_I , i.e., the secondary electrons generated by the incident beam upon entering the sample. Secondary electrons generated by backscattered electrons when leaving the sample are termed SE_{II} . Secondary electrons generated when the backscattered electrons strike a lens pole piece or the sample chamber's wall and by primary electrons hitting the aperture are, respectively, called SE_{III} and SE_{IV} . Although SE_{III}



Scanning Electron Microscopy, Fig. 3 SEM micrographs of a polymer membrane that consists of electrospun fibrous scaffold for water filtration. The fibrous scaffold has an average diameter of 200 nm. The high-resolution SEM image on the *right* shows that cellulose nanofibers (5–10 nm in diameter) infused into the scaffold form a

cellulose network to enhance the membrane's ability to remove bacteria and viruses. The images were taken with JEOL7600 SEM at operation voltage of 0.5 eV. A thin layer of carbon was coated on the sample to avoid charging. Note the bright contrast at the edge of the fibers generates a topological view of the fibrous structure

contain the information on the sample, SE_{IV} do not. Furthermore, there are fast secondary electrons, with energy higher than 50 eV. Bias experiments, wherein a positive dc voltage is applied to the sample to suppress the emission of the secondary electrons, were mainly designed to separate the SE_I from backscattered electrons; they cannot distinguish SE_I from SE_{II} , which for a very thin specimen should be negligible. Measuring other types of secondary electrons, including those with high energy, would require a different bias experiment. Heavy elements (high atomic number) that backscatter electrons more strongly than do light elements (low atomic number), and thus appear brighter in the image, can be used to yield contrast that contains information of a sample's chemical composition.

Figure 3 shows SEM images of a nanofiber containing polymer membrane developed for water filtration with enhanced ability to remove bacteria and viruses. The image on the right shows the structural network formed by the 5–10 nm diameter cellulose nanofibers. It reveals astonishing topographic details on how the cellulose nanofibers are interwoven with scaffold. In SEM images, flat surfaces give even contrast, while curved surfaces and sharp edges often appear brighter (high image intensity). This is due to the increased escape of secondary electrons from the

top and side surfaces when the interaction volume intercepts them. Since the secondary-electron detector is biased with a high acceleration voltage, a surface facing away from the detector still can be imaged. Surfaces tilted away from the normal to the beam allow more secondary electrons to escape [13].

Atomic Imaging Using Secondary Electrons

In the last decade or so, high-resolution SEM has proven an indispensable critical dimension metrology tool for the semiconductor industry. The road map for semiconductor nanotechnology identifies the need for ultrahigh-resolution SEM in the quest for ever-decreasing device sizes. In a SEM, the size of the imaging probe often determines the instrument's resolution power. The probe size d (measured in full width and half maximum) is a function of the beam convergence; half-angle α is an incoherent sum of contributions from source size, diffraction limit, spherical aberration, and chromatic aberration and is given by

$$d^2 = \left(\frac{4i_p}{\beta\pi^2\alpha^2} \right)^2 + \left(\frac{0.6\lambda}{\alpha} \right)^2 + (0.5C_s\alpha^3)^2 + \left(C_c\alpha \frac{\Delta E}{E} \right)^2$$

where i_p is the probe current, α is the convergence half-angle, β is the source brightness, λ is the electron wavelength at beam energy E , ΔE is the energy spread, and C_s and C_c are, respectively, the spherical and chromatic aberration coefficients of the probe-forming lens. The first term on the right side of the equation is the probe size defined by the source size, the second term is due to the diffraction limit, and the third and the fourth terms are due, respectively, to the spherical and chromatic aberrations. Recent advancement on correcting spherical aberration in SEM and STEM diminishes C_s to zero, thus eliminating the third term and produces a small probe. It is important to note that the probe size also depends on the energy spread in the fourth term that includes the energy distribution of the primary electron beam and the fluctuation of the instruments' acceleration voltage.

In SEM, the image resolution depends not only on the instrument but also on the sample, or, more accurately, on the sampling volume of the sample from which the signal is generated. When the incident electrons impinge on a point of the sample's surface, they interact with atoms in the sample via elastic (change trajectory) and inelastic (lose energy) interactions. The size of the interaction volume depends on the electron's landing energy, the atomic number of the sample, and its density. The simultaneous energy loss and change in trajectory spread beam into the bulk of the sample and produce an interaction volume therein that, for a thick sample, can extend from less than 100 nm to around 5 μm into the surface, i.e., more than an order of magnitude larger than the original probe size. For a very thin sample (a few nm thick), the interaction volume, as the first approximation, might be defined as the probe size.

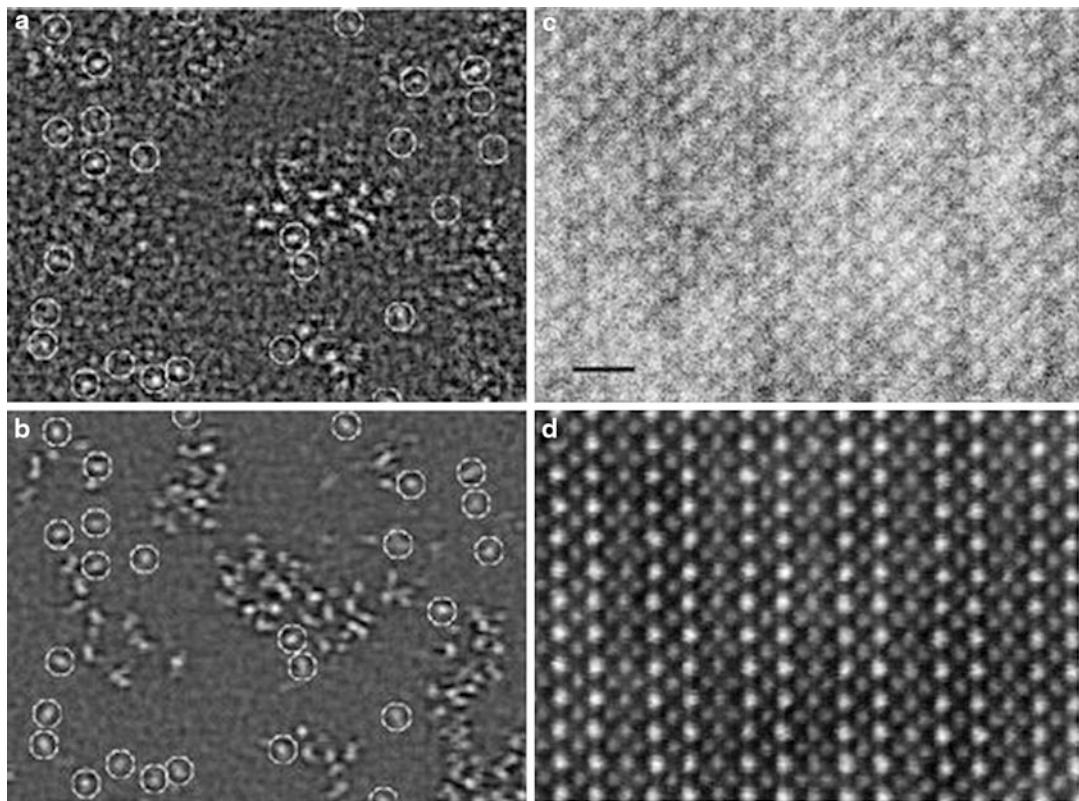
Figure 4 illustrates the atomic resolution images using secondary electrons on single uranium atoms (a) and the (010) surface of a $\text{YBa}_2\text{Cu}_3\text{O}_7$ crystal (c) recorded on the Hitachi HD2700C SEM/STEM (Fig. 2d). For comparison, the corresponding annular dark-field images using transmitted electrons, (b) and (d), respectively, are included. The equally sharp images in SE and ADF suggest negligible imaging delocalization. Such an attainable resolution was

attributed to the combination of several factors: better design of the electro-optics of the instrument (including ultrahigh electric and mechanical stabilities) and the detector, aberration correction that reduces the probe size and increases the probe current, and the higher operation voltage that beneficially assures a very small volume of beam interaction for a thin sample.

The physical mechanism of producing the low-energy secondary electrons traditionally is ascribed to inelastic scattering and decay of collective electron excitation with the incident electron giving up, say, 20 eV, to produce a secondary electron with energy 20 eV energy minus the work function of the surface. Since the momentum transfer (scattering angle) of the scattering with 20 eV energy loss is small, the transfer should be delocalized to an area $>1 \text{ nm}$; thus, atomic imaging using secondary electrons was not considered possible. Recent studies suggest that this is not the primary mechanism for secondary-electron imaging at least on a thin sample [11]. The secondary electrons responsible for atomic-scale resolution are generated by inelastic scattering events with large momentum transfer, including those from inner shell orbitals, which give rise to a sharp central peak in the point spread function for signal generation.

In general, four steps are involved in producing the signal that is used to form a secondary-electron (SE_1) image [12]: (1) the generation of secondary electrons through the inelastic scattering of primary electrons in the sample, at a generation rate, G ; (2) random motion of these secondary electrons, which are scattered by atoms of the specimen both elastically and inelastically (potentially creating other secondary electrons of lower energies), such that, on average, T electrons reach the sample surface for each secondary electron generated; (3) the escape of secondary electrons over the potential barrier at the sample surface of the specimen, with an average probability P ; and (4) the acceleration of the emitted electrons in vacuum, such that a fraction D reaches the electron detector.

The secondary-electron signal S is a product of these four factors: $S = G \cdot T \cdot P \cdot D$. To generate contrast in a scanned probe image, one or more of the



Scanning Electron Microscopy, Fig. 4 (a–b) Simultaneous atomic imaging using secondary electrons (secondary-electron mode, a) and transmitted electrons (annual dark-field mode, b) of uranium individual atoms on a carbon support (raw data). The circles mark the single uranium atoms. The atoms shown in (b) but not in (a) are

presumably those on the back side of the support. (c–d) Simultaneous atomic imaging using secondary electrons (c) and transmitted electrons (d) of $\text{YBa}_2\text{Cu}_3\text{O}_7$ superconductor viewing along the [010] direction (raw data). The scale bar in (c) is 0.7 nm

above steps must depend on the x -coordinate of the electron probe in the scan direction, i.e.,

$$\begin{aligned} \frac{dS}{dx} = & I_0 TPD \frac{dG}{dx} + I_0 GPD \frac{dT}{dx} + I_0 GTD \frac{dP}{dx} \\ & + I_0 GTP \frac{dD}{dx} \end{aligned}$$

For most non-atomically resolved SE images obtained in a SEM, $\frac{dT}{dx}$ provides the main contrast mechanism: Secondary electrons created at an inclined surface or close to a surface step have an increased probability of escape, resulting in surface-topography contrast [14]. Less commonly, variations in surface work function contribute additional contrast by providing a

nonzero $\frac{dP}{dx}$. In voltage-contrast applications, changes in surface voltage provide a nonzero $\frac{dD}{dx}$. Atomic number contrast is possible if the specimen is chemically inhomogeneous and G varies with atomic number, yielding a nonzero $\frac{dG}{dx}$. However, for atomic imaging in a thin sample (Fig. 4), the dominant mechanism can be quite different. $\frac{dG}{dx}$ is likely to play an important role in the atomic-scale contrast as a consequence of Z-dependence inelastic scattering cross section and channeling effect for crystals (for thickness in the order of extinction distance, it should be minor). Because secondary electrons are generated through inelastic scattering of the incident electrons, $\frac{dG}{dx}$ is limited by the delocalization of the scattering process, which may be described by the point

spread function for inelastic scattering. The term $\frac{dT}{dx}$ should be small, for adatoms or surface atoms that lie on the detector side of the sample. $\frac{dP}{dx}$ would not become important unless the effective work function varies on an atomic scale, and $\frac{dD}{dx}$ must be also negligible at atomic scale. These assumptions are reasonable because the scattering process disperses secondary electrons over a range of x that is comparable to the escape depth, typically 1–2 nm. Consequently, T , P , and D are x -averages that vary little with x on an atomic scale. For uranium atoms on a carbon substrate, the argument is even simpler; these atoms lie outside the solid, so the terms T and P are not applicable.

Future Remarks

Secondary-electron imaging is the most popular mode of operation of the scanning electron microscope (SEM) and traditionally is used to reveal surface topography. Nevertheless, this imaging method never was regarded as being on the cutting edge of performance, due to its perceived limited spatial resolution in comparison with its TEM or STEM counterparts using transmitted electrons. Recent work using aberration-corrected electron microscopes demonstrated that secondary-electron signals in the SEM can resolve both crystal lattices and individual atoms, showing SEM's unprecedented and previously unrealized imaging capabilities. Furthermore, the work demonstrates the incompleteness of present understanding of the formation of secondary-electron images. Secondary-electron imaging using high acceleration voltage with thin samples can now compete with TEM on spatial resolution and provide new capabilities, such as depth-resolved profiles, at the atomic level. There seems to be no fundamental reason why atomic resolution in secondary-electron imaging could not be achieved at the accelerating voltages of 0.5–40 keV that are currently used in conventional SEMs. It remains to be seen whether the integrated spherical and chromatic aberration of a probe-forming objective lens can be corrected to a sufficient degree at low operation voltages.

One clear outcome thus far is the importance of preparing samples with clean surfaces in order to obtain interpretable and reproducible results.

Cross-References

- ▶ [Robot-Based Automation on the Nanoscale](#)
- ▶ [Scanning Tunneling Microscopy](#)
- ▶ [Transmission Electron Microscopy](#)

References

1. Knoll, M.: Aufladepotential und sekundäremission elektronenbestrahlter körper. *Z. Tech. Phys.* **16**, 467–475 (1935)
2. von Ardenne, M.: Das Elektronen-Rastermikroskop. Praktische Ausführung. *Z. Tech. Phys.* **19**, 407–416 (1938) (in German)
3. Wells, O.C., Joy, D.C.: The early history and future of the SEM. *Surf. Interface Anal.* **38**, 1738–1742 (2006)
4. Oatley, C.W.: The early history of the scanning electron microscope. *J. Appl. Phys.* **53**, R1–R13 (1982)
5. Everhart, T.E., Thornley, R.F.M.: Wide-band detector for micro-microampere low-energy electron currents. *J. Sci. Instrum.* **37**, 246–248 (1960)
6. Goldstein, G.I., Newbury, D.E., Echlin, P., Joy, D.C., Fiori, C., Lifshin, E.: *Scanning Electron Microscopy and X-ray Microanalysis*. Plenum, New York (1981)
7. Pawley, J.: The development of field-emission scanning electron microscopy for imaging biological surfaces. *Scanning* **19**, 324–336 (1997)
8. Tanaka, K., Mitsushima, A., Kashima, Y., Osatake, H.: A new high resolution scanning electron microscope and its application to biological materials. In: *Proceedings of the 11th International Congress on Electron Microscopy*, Kyoto, pp. 2097–2100 (1986)
9. Liu, J., Cowley, J.M.: High resolution SEM in a STEM instrument. *Scanning Microsc.* **2**, 65–81 (1988)
10. Howie, A.: Recent developments in secondary electron imaging. *J. Microsc.* **180**, 192–203 (1995)
11. Zhu, Y., Inada, H., Nakamura, K., Wall, J.: Imaging single atoms using secondary electrons with an aberration-corrected electron microscope. *Nat. Mater.* **8**, 808–812 (2009)
12. Inada, H., Su, D., Egerton, R.F., Konno, M., Wu, L., Ciston, J., Wall, J., Zhu, Y.: Atomic imaging using secondary electrons in a scanning transmission electron microscope: experimental observations and possible mechanisms. *Ultramicroscopy* **111**(7), 865–876 (2011). doi:10.1016/j.ultramic.2010.10.002. Invited articles for the special issue in honor of John Spence
13. Joy, D.C.: Beam interactions, contrast, and resolution in the SEM. *J. Microsc.* **136**, 241–258 (1984)
14. Reimer, L.: *Scanning Electron Microscopy*, 2nd edn. Springer, New York (1998)

Scanning Electron Microscopy and Secondary-Electron Imaging Microscopy

- ▶ [Scanning Electron Microscopy](#)
-

Scanning Force Microscopy in Liquids

- ▶ [AFM in Liquids](#)
-

Scanning Kelvin Probe Force Microscopy

- ▶ [Kelvin Probe Force Microscopy](#)
-

Scanning Near-Field Optical Microscopy

Achim Hartschuh
Department Chemie and CeNS,
Ludwig-Maximilians-Universität München,
Munich, Germany

Synonyms

[Near-field scanning optical microscopy \(NSOM\)](#)

Definition

Scanning near-field optical microscopy (SNOM) is a microscopic technique for nanostructure investigation that achieves sub-wavelength spatial resolution by exploiting short-ranged interactions between a sharply pointed probe and the sample mediated by evanescent waves. In general, the resolution of SNOM is determined by the lateral probe dimensions and the probe-sample distance. Images are obtained by raster-scanning the probe

with respect to the sample surface corresponding to other scanning-probe techniques. As in conventional optical microscopy, the contrast mechanism can be combined with a broad range of spectroscopic techniques to study different sample properties, such as chemical structure and composition, local stress, electromagnetic field distributions, and the dynamics of excited states.

Introduction

Optical microscopy forms the basis of most of the natural sciences. In particular, life sciences have benefited from the fascinating possibility to study smallest structures and processes in living cells and tissue. Optical techniques feature extremely high detection sensitivity reaching single molecule sensitivity in fluorescence, Raman scattering and absorption spectroscopy. Besides the direct visualization, chemically specific information is obtained through Raman spectroscopy.

The resolution of conventional optical microscopes, however, is limited by diffraction, a consequence of the wave nature of light, to about half the wavelength. Concepts extending optical microscopy down to nanometer length scales below the diffraction-limit are distinguished into far-field and near-field techniques. Far-field techniques rely on the detection of propagating waves at distances from the source larger than the wavelength, while near-field techniques exploit short ranged evanescent waves.

Scanning near-field optical microscopy, initiated by pioneering work of Pohl, Lewis and others in the late 1980s and the beginning of 1990s gave access to nanoscale resolution for the first time. The history of near-field optics is reviewed in [1]. In addition numerous review articles and books exist describing fundamentals and applications (see e.g., [2–4]).

This entry introduces first the key physical principles beginning with the role of evanescent and propagating waves, and the loss of spatial information upon light propagation. Concepts of near-field detection are shown using different pointed probes. The next sections describe the experimental realization and present several

applications of SNOM. The outlook addresses future prospects of SNOM and remaining challenges.

Key Principles and Concepts

Near-field optics has its origin in the effort of overcoming the diffraction limit of optical imaging. The physical origin of this limit is sketched in the following starting with the distinction between propagating waves that form the optical far-field of a radiation source and their evanescent counterpart that dominate the optical near-field. A powerful tool to describe wave propagation is the so-called angular spectrum representation of fields expressing the electromagnetic field E in the detector plane at z as the superposition of harmonic plane waves of the form $\exp(i\vec{k}\cdot\vec{r} - i\omega t)$ with amplitudes $\bar{E}(k_x, k_y, z = 0)$ emanating from the source plane at $z = 0$ [2].

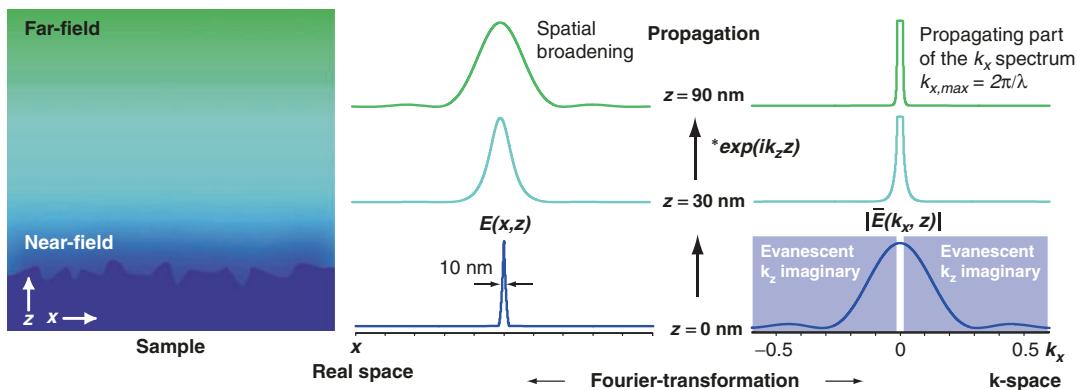
$$E(x, y, z) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \bar{E}(k_x, k_y, z = 0) e^{i(k_x x + k_y y)} e^{\pm ik_z z} dk_x dk_y \quad (1)$$

The wave vector \vec{k} describing the propagation direction of the wave is represented by its components $\vec{k} = (k_x, k_y, k_z)$ while its length is fixed by the wavelength of light λ and the refractive index of the medium n through $|\vec{k}| = \sqrt{k_x^2 + k_y^2 + k_z^2} = 2\pi n/\lambda$. In Eq. 1, the time dependence of the fields has been omitted for clarity. For simplicity the following discussion is limited to the x - z -plane and $n = 1$ such that $k_z = \sqrt{4\pi^2/\lambda^2 - k_x^2}$. In Eq. 1 the term $e^{\pm ik_z z}$ controls the propagation of the associated wave: For $k_x \leq 2\pi/\lambda$ the component k_z is real and the corresponding wave with amplitude $\bar{E}(k_x, z = 0)$ propagates along the z -axis oscillating with $e^{-ik_z z}$. If $k_x > 2\pi/\lambda$ the component k_z becomes complex and $e^{-|k_z|z}$ describes an exponential decay of the associated wave that is therefore evanescent. As a result, only waves with $k_x \leq 2\pi/\lambda$ can propagate and contribute to the field far from the source

forming the far-field. Figure 1 schematically illustrates this behavior: In the center, the electric field $E(x, z)$ emanating from a narrow sub-wavelength source at $z = 0$ is shown together with its angular spectrum $\bar{E}(k_x, z = 0)$ calculated by the inverse of Eq. 1. The wave amplitudes \bar{E} result from the Fourier-transformation of E with respect to the spatial coordinate x . As for the correlation between time and frequency domain, where a short optical pulse requires a broad frequency spectrum, a sharp field distribution requires a broad spectrum of spatial frequencies k_x .

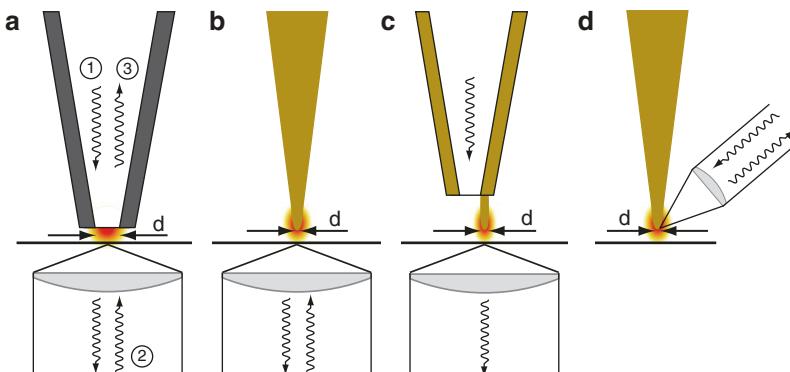
Since only waves with limited spatial frequencies can propagate, the spectral width rapidly decreases with increasing distance from the source z leading to fast broadening of the electric field distribution in real space. In other words, propagation corresponds to low-pass filtering with frequency limit $k_{x,\max} = 2\pi/\lambda$. The far-field thus contains limited spatial frequencies equivalent to limited spatial information. To overcome this limitation different near-field concepts have been developed that are outlined in the following.

The key concept of SNOM is the probing of the sample near-field that contains the evanescent waves using a sharply pointed probe. Since evanescent waves decay rapidly for increasing distance to the source, the probe needs to be in close proximity to the sample. Waves with large k_x components that carry high spatial information decay most rapidly following $e^{-|k_z|z}$ as can be seen from Fig. 1. Hence the spatial resolution obtained in an SNOM experiment drops fast with increasing tip-sample distance z . As for other scanning probe techniques that exploit short-ranged interactions, such as atomic force and scanning tunneling microscopy, AFM and STM, respectively, the lateral resolution is also determined by the lateral dimension of the probe. Two conceptually different types of probes can be distinguished: The first confines and samples electromagnetic fields using an aperture with sub-wavelength diameter (Fig. 2a). The second exploits the antenna concept that couples locally enhanced near-fields to propagating waves and vice versa (Fig. 2b-d). The two types, termed aperture and antenna probe, respectively, are illustrated in the following.



Scanning Near-Field Optical Microscopy,
Fig. 1 Scheme illustrating the propagation of waves and the loss of spatial information. Initial field distribution $E(x, z = 0)$ at a 10-nm wide source in the x - z -plane (center) and corresponding angular spectrum $\bar{E}(k_x, z = 0)$ (right). Near the source the spectrum contains both evanescent and propagating waves. Upper panels illustrate the evolution

of the fields at $z = 30$ nm and $z = 90$ nm distance for a source wavelength of $\lambda = 500$ nm in vacuum. Only waves with $k_x \leq 2\pi/\lambda$ propagate. Evanescent waves decay exponentially following $e^{-|k_z|z}$. The decay of high spatial frequencies leads to spatial broadening and loss of spatial information in the far-field



Scanning Near-Field Optical Microscopy,
Fig. 2 Schematics of the most common SNOM probes and configurations: (a) Aperture probes are utilized in different modes: Near-field excitation – far-field collection ①, far-field excitation – near-field collection ② and near-field excitation – collection mode ③. (b) Tip-enhanced near-field optical microscopy (TENOM) using local field enhancement at a sharp metal antenna probe upon far-field

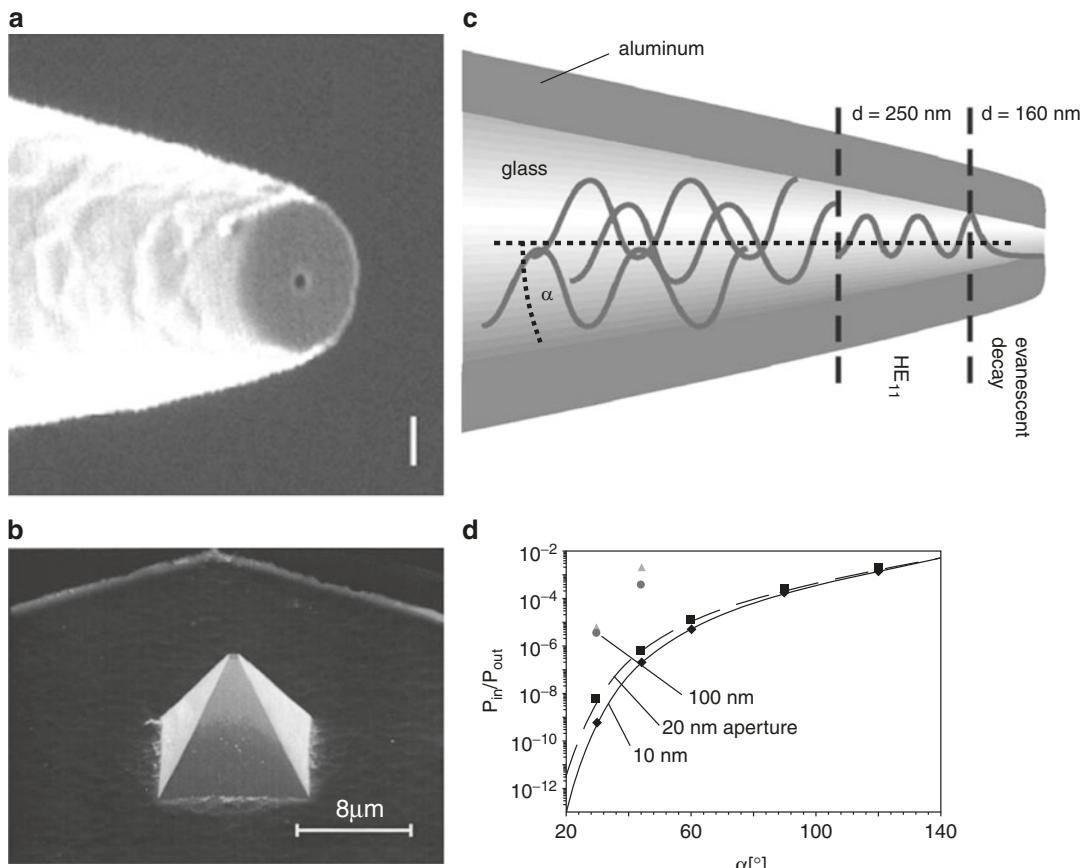
excitation. (c) Tip-on-aperture (TOA) probe using local field enhancement at an antenna probe that is excited in the near-field of an aperture probe. (d) Scattering-SNOM (s-SNOM) based on far-field illumination and detection of local scattering at a sharp antenna probe. The label d indicates the structural parameter that determines the achievable spatial resolution

Aperture Probes

Aperture probes confine light by squeezing it through a sub-wavelength hole (Fig. 2a). This approach, termed aperture-SNOM, provides an enormous flexibility regarding signal formation. Different operation modes can be used that are capable of local sample excitation and/or local light collection. Depending on which step of the

experiment exploits near-field interactions to obtain sub-wavelength resolution, aperture-SNOM can be implemented in excitation ①, collection ② and excitation-collection ③ mode (Fig. 2a).

The original scheme was proposed by Synge in 1928. He suggested to use a strong light source behind a thin, opaque metal film with a 100-nm



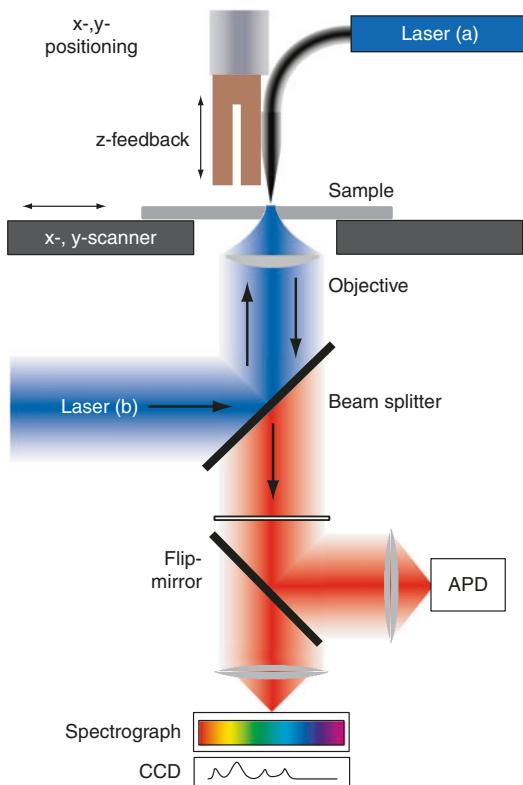
Scanning Near-Field Optical Microscopy, Fig. 3 (a) Scanning-electron microscopy (*SEM*) image of an aperture probe formed by a metal-coated tapered fiber with an aperture diameter of 70 nm (scale bar 200 nm) (Reprinted with permission from Veerman et al. [5]. Copyright 1999, John Wiley and Sons) (b) *SEM* image of a metallic hollow aperture probe microfabricated on a Si cantilever. The aperture diameter is about 130 nm (Reprinted with permission from Mihalcea et al. [6]. Copyright 1996, American Institute of Physics) (c) Schematic of the mode propagation

in a tapered aperture probe. For probe diameters below the cut-off diameter, here $d = 160$ nm, the intensity decays exponentially toward the aperture. (d) Transmission of tapered probes determined as the ratio of input versus output power $P_{\text{in}}/P_{\text{out}}$ as a function of the cone angle α defined in (c). For smaller cone angles, the distance between cut-off and aperture increases leading to extremely low transmission (c and d) (Reprinted with permission from Hecht et al. [7]. Copyright 2000, American Institute of Physics)

diameter hole in it as a very small light source. In 1984, two groups adopted this scheme and presented the first experimental realizations in the optical regime [1]. The aperture was formed at the apex of a sharply pointed transparent probe tip coated with metal. Raster-scanning the probe was made possible by the scanning technology developed in the context of scanning tunneling microscopy (STM). In Fig. 3a, a scanning electron microscopy (SEM) image of an aperture probe consisting of a metal-coated tapered glass fiber is

shown. At the front surface a well-defined aperture with diameter of 70 nm is seen.

Analytical expressions for the electric field distribution in a sub-wavelength aperture in a metallic screen were already presented in 1944 and 1950 by Bethe and Bouwkamp. Numerical simulations for metal-coated tapered fiber probes show that strong fields pointing in axial direction occur at the rim due to local field-enhancement by the metal coating (see Fig. 4). The center of the aperture is dominated by a weaker horizontal



Scanning Near-Field Optical Microscopy, Fig. 4 Schematic of an experimental setup applicable to aperture-probe SNOM (laser (a), excitation mode) and TENOM (laser (b)) in case of transparent samples. In the case of TENOM the fiber probe would be replaced by an optical antenna, for example, an etched metal wire. The probe tip is positioned in the focus of the microscope objective by piezo-electric actuators. The sample is raster-scanned using a closed-loop x-y-scanner while both laser and probe position remain fixed. The tip-sample distance is controlled using a tuning-fork shear-force feedback scheme. The optical signal is collected by the objective and detected either by a highly sensitive avalanche photodiode (APD) or energetically resolved using a spectrograph and a CCD

component [2]. The optical field distribution can be determined experimentally by raster-scanning single fluorescent molecules that act as point-like dipoles across the aperture while recording the fluorescence intensity [5] (see section “Fluorescence Microscopy” and Fig. 3).

Tapered aperture probes suffer from low light transmission due to the cut-off of propagating wave-guide modes. For probe diameters below the cut-off diameter only evanescent waves

remain and the intensity decays exponentially toward the aperture (Fig. 2c, d). Probe designs, therefore, aim at maximizing the cone angle that determines the distance between aperture and cut-off diameter. Hollow-cantilever probes feature relatively large cone-angles as compared to fiber-based probes (Fig. 2b). On the other hand, the input power needs to be limited because of the damage threshold of the metal coating in case of fiber-based probes. Due to the limited transmission and the skin-depth of the optical fields on the order of several tens of nanometers, most aperture-SNOM measurements are carried out with apertures of 50–100 nm.

Antenna Probes

Antenna probes act as transmitter and receiver coupling locally enhanced near-fields to propagating waves and vice versa (Fig. 2b–d) [8]. To distinguish this approach from the earlier implementations based on apertures, it is also termed apertureless-SNOM or a-SNOM. Antenna probes can be used in two different techniques: (1) Scattering type microscopy [9, 10], also termed scattering-SNOM or s-SNOM, in which the tip locally perturbs the fields near a sample surface. The response to this perturbation is detected in the far-field at the frequency of the incident light corresponding to elastic scattering (Fig. 2d). (2) Tip-enhanced near-field optical microscopy (TENOM) in which locally enhanced fields at laser-illuminated metal structures are used to increase the spectroscopic response of the system at frequencies different from that of the incident light [8, 11] (Fig. 2b, c). The flexibility of this technique allows the study of a variety of spectroscopic signals including Raman scattering (tip-enhanced Raman spectroscopy (TERS)), and fluorescence as well as time-resolved measurements. In the following, the signal formation in the case of optical antennas is sketched.

Elastic scattering signal. The near-field interaction can be treated within a simplified model in which the tip is replaced by a polarizable sphere. Due to the antenna properties of the tip, laser excitation with incident field E_i creates a dominating dipole oriented along the tip axis in z-direction normal to the sample surface. This dipole induces

a mirror dipole in the sample depending on its dielectric properties. The mirror dipole's field, decreasing with the third power of distance, interacts with the tip dipole. Solving the system of electrostatic equations that describes the multiple interaction between tip and mirror dipoles neglecting retardation yields an effective polarizability of the coupled tip-sample system which fully expresses the influence of the sample.

$$\alpha_{\text{eff}} = \left(\frac{(\alpha(1 + \beta)/1 - (\alpha\beta))}{(16\pi(\alpha + z)^3)} \right) \quad (2)$$

The scattered field can then be calculated from $E_s \propto \alpha_{\text{eff}} E_i$ reflecting the short-ranged interaction required for sub-diffraction resolution. Since the laser illuminates a greater part of the tip and the sample, elimination of background-scattering contributions is crucial. Efficient background suppression can be achieved by demodulating the detected intensity signal at higher harmonics of the tapping mode frequency (see section “[Instrumentation](#)”). Besides the amplitude of the scattered field, its phase can be retrieved using interferometric heterodyne detection [9].

Raman scattering signal. In the case of Raman scattering, the total signal depends on the product of the excitation and emission rates $k^{\text{ex}}(\lambda^{\text{ex}}) k^{\text{rad}}(\lambda^{\text{rad}})$. As a consequence, the total signal enhancement scales with the fourth power of the field enhancement for small differences between the excitation λ^{ex} and emission wavelength λ^{rad} and assuming that the field enhancement at the tip does not depend sensitively on the wavelength.

$$M_{\text{Raman}} = \frac{k_{\text{tip}}^{\text{ex}}}{k_0^{\text{ex}}} \cdot \frac{k_{\text{tip}}^{\text{rad}}}{k_0^{\text{rad}}} \approx f^4 \quad (3)$$

The factor f measures the ratio between tip-enhanced E_{tip} and non-enhanced electric field E_0 in the absence of the tip. For the general case of surface-enhanced Raman scattering (SERS), Raman enhancement factors are reported reaching up to 12 orders of magnitude for particular multiple particle configurations involving interstitial sites between particles or outside sharp surface protrusions. Since the signal scales with the fourth

power, already moderate field enhancement, predicted for a single spherical particle to be in the range of $f = 10\text{--}100$ is sufficient for substantial signal enhancement.

Fluorescence signal. The fluorescence intensity depends on the excitation rate k^{ex} and the quantum yield η denoting the fraction of transitions from excited state to ground state that give rise to an emitted photon. The quantum yield is expressed in terms of the radiative rate k^{rad} and the non-radiative rate k^{nonrad} through $\eta = k^{\text{rad}}/(k^{\text{rad}} + k^{\text{nonrad}})$. Accordingly, the fluorescence enhancement due to the presence of the metal tip can be written as

$$M_{\text{Flu}} = \left(\frac{E_{\text{tip}}}{E_0} \right)^2 \left(\frac{\eta_{\text{tip}}}{\eta_0} \right) = f^2 \left(\frac{\eta_{\text{tip}}}{\eta_0} \right). \quad (4)$$

Here, it is assumed that the system is excited far from saturation. From Eq. 4 it is clear that TENOM works most efficiently for samples with small fluorescence quantum yield η_0 such as semiconducting single-walled carbon nanotubes [11]. For highly fluorescent samples such as dye molecules, the quantum yield η_0 is already close to unity and cannot be enhanced further.

Because of the small separation between emitter and metal tip required for high spatial resolution, non-radiative transfer of energy from the electronically excited state to the metal followed by non-radiative dissipation in the metal has to be taken into account. This process represents an additional competing non-radiative relaxation channel and reduces the number of detected fluorescence photons. Metal-induced fluorescence quenching can also be exploited for image contrast formation (see e.g., [12]). In this case M_{Flu} in Eq. 4 becomes smaller than unity.

While the theory of energy transfer between molecules and flat metal interfaces is well understood in the framework of phenomenological classical theory, nanometer-sized objects are more difficult to quantify [2]. Tip- and particle-induced radiative rate enhancement and quenching has been studied in literature both experimentally and theoretically. Experiments on model systems formed by single dipole emitters such as molecules and semiconductor nanocrystals and

spherical metal particles revealed a distance-dependent interplay between competing enhancement and quenching processes. While semiconducting tips cause less efficient quenching, they also provided weaker enhancement because of their lower conductivity at optical frequencies.

Polarization and angular-resolved detection of the fluorescence signal of single emitters demonstrated that the fluorescence rate enhancement provided by the optical antenna also results in a spatial redistribution of the emission [13]. The same redistribution can be expected to occur for tip-enhanced Raman scattering. The spatial distribution of the enhanced electric field follows approximately the outer dimensions of the tip apex. Since the signal enhancement scales with higher orders of field enhancement, the optical resolution can surpass the size of the tip [11]. Stronger fields and field confinement are observed for so-called gap modes formed by metal tips on top of metal substrates.

The scheme depicted in Fig. 2b shows that in addition to the signal resulting from the near-field tip–sample interaction the confocal far-field signal contribution is detected, representing a background. This background originates from a diffraction-limited sample volume that is far larger than the volume probed in the near-field. The near-field signal has to compete with this background, and strong enhancement is required to obtain clear image contrast. This requirement is relaxed in case of low-dimensional sample structures such as spatially isolated molecules or one-dimensional nanostructures [11]. The near-field signal to background ratio can be improved by exploiting the non-linear optical response of sample and tip. Examples include two-photon excitation of fluorescence using a metal tip antenna and the application of the four-wave mixing signal of a metal particle dimer as local excitation source.

Instrumentation

Near-field optical microscopy exploits short-ranged near-field interactions between sample and probe. SNOM instruments thus require a

mechanism for tip-sample distance control working on the scale of nanometers. Typical implementations utilize other non-optical short-ranged probe–sample interactions such as force or tunneling current used for topography measurements in AFM and STM, respectively. During image acquisition by raster-scanning the sample with respect to the tip, optical and topographic data are thus obtained simultaneously. Due to the strong tip-sample distance dependence of the near-field signal, cross-talk from topographic variations is possible that can lead to artifacts in the optical contrast. Tip-sample distance curves need to be measured to prove unequivocally the near-field origin of the observed image contrast. Since the optical signal results from a single sample spot only, SNOM instruments are often based on a confocal laser scanning optical microscope equipped with sensitive photodetectors.

Scattering-SNOM is often implemented with intermittent-contact-mode AFM in which the tip-sample distance is modulated sinusoidally at frequencies typically in the range of 10–500 kHz. The elastically scattered laser light intensity is demodulated by the tapping-mode frequency using lock-in detection for background suppression.

Since Raman and fluorescence signals are typically weak requiring longer acquisition times, signal demodulation is more challenging. In the case of single-photon counting time-tagging can be used to retrieve the signal phase with respect to the tapping oscillation. While this can be applied to single-color experiments, spectrum acquisition using CCD cameras in combination with spectrometers is not feasible at typical tapping frequencies.

In most aperture-SNOM and TENOM experiments, the tip-sample distance is kept constant by either using STM, contact/non-contact AFM, or shear-force feedback. Sensitive piezoelectric tuning-fork detection schemes operating at small interaction forces have been developed and are used for fragile fiber and antenna probes [2].

Probe Fabrication

The near-field probe forms the crucial part of an SNOM setup since both optical and topographic

signals are determined by its short-ranged interactions with the sample. Instrument development, therefore, focuses mainly on the design of optimized tip concepts and tip geometries as well as on fabrication procedures for sharp and well-defined probes with high reproducibility. These continuing efforts benefit substantially from improving capabilities regarding nanostructuring and nanocharacterization, using, for example, focused ion beam milling (FIB) or electron beam lithography (EBL).

Aperture Probes

Optical fiber probes. Sharply pointed optical fiber probes were the first to provide sub-diffraction spatial resolution and are widely used as nanoscale light sources, light collectors or scatterers (Figs. 2a and 3a). Probe fabrication requires a number of steps starting with the formation of a tapered optical fiber. Typically two different methods are used. Chemical etching of a bare glass fiber dipped into hydro-fluoric (HF) acid yields sharp tips [4, 7]. The surface tension of the liquid forms a meniscus at the interface between air, glass, and acid. A taper is formed due to the variation of the contact angle at the meniscus, while the fiber is etched and its diameter decreased. Chemical etching allows reproducible production of larger quantities of probes in a single step. A specific advantage is that the taper angle can be tuned and optical probes with correspondingly large transmission coefficient can be produced.

The second method combines local heating using a CO₂ laser or a filament and subsequent pulling until the fiber is split apart. The resulting tip shapes depend heavily on the temperature and the timing of the heating and pulling, as well as on the dimensions of the heated area. The pulling method has the advantage of producing tapers with very smooth surfaces, which positively influences the quality of the evaporated metal layer. Etched probes, on the other hand, typically feature rough surfaces. Pulled fibers, however, have small cone angles and thus reduced optical transmission as well as flat end-faces limiting the minimum aperture size.

The aperture is formed during the evaporation of aluminum. Since the evaporation takes place under an angle slightly from behind, the deposition rate of metal at the apex is much smaller than on the sides. This geometrical shadowing effect leads to the self-aligned formation of an aperture at the apex.

The ideal aperture probe should have a perfectly flat end face to position a sample as close as possible into the near-field of the aperture. Conventional probes generally have a roughness determined by the grain size of the aluminum coating, which is around 20 nm at best. Due to the corrugated end face the distance between aperture and sample increases, which lowers the optical resolution and decreases the light intensity on the sample. Furthermore, these grains often obscure the aperture, which makes the probe ill-defined and not suited for quantitative measurements [4, 5, 7]. Subsequent focused ion beam (FIB) milling can be used to form high definition SNOM probes with well-defined end face (Fig. 3).

Microfabricated cantilevered probes. Hollow-pyramid cantilevered probes can be batch-fabricated with large taper angles. Lithographic patterning of an oxidized silicon wafer first defines the position of the aperture and the dimensions of the cantilever beam by structuring the oxide layer [6]. Anisotropic etching of the exposed silicon with buffered HF forms a pyramidal groove for the tip and trenches for the cantilever beam. After removing the oxide layer on the opposite side anisotropic etching is used to open a small aperture in the pyramidal groove. A 120-nm chromium layer is deposited on the back side forming the hollow pyramidal tip that is finally freed by isotropic reactive ion etching.

Si₃N₄ tips can be fabricated by dry etching in a CF₄ plasma and covering with thin aluminum films. Microfabricated probes based on quartz tips attached to silicon cantilevers have also been reported [2, 7]. First sharp quartz tips were produced in hydrofluoric acid followed by coating with thin films of aluminum and silicon nitride. Reactive ion etching was then used to selectively remove the silicon nitride from the tip apex, while the remaining film served as a mask for

wet-etching of the protruding aluminum in a standard Al-etching solution, leaving a small aperture on the apex of the tip. For light coupling, windows are etched into the backside of the levers at the position of the tip.

Microfabricated tips provide several advantages over fiber-based probes. The mechanical stability is typically increased and often sufficient to measure also in contact AFM mode without destroying the tip. A reproducible fabrication of the tips leads to a well-defined aperture shape. Large taper angles of around 70° shift the position of the mode cut-off closer to the aperture, resulting in higher transmission.

Antenna Probes

Sharp metal tip are fabricated in a single step by simple electrochemical etching. In the case of gold, pulsed or continuous etching in hydrochloric acid (HCL) or a mixture of HCL and ethanol routinely yields tips with diameters in the range of 30–50 nm. These tips can either be used in shear-force mode after gluing to the prong of a tuning-fork or in STM-mode. Silicon or silicon nitride cantilevered probes can be coated by a thin metal film through evaporation. Subsequent nanostructuring by FIB-milling can be used to optimize tip parameters.

Tip-on-aperture probes need to be fabricated in a series of sequential steps [14] (Fig. 2c). First, a fiber-based aperture probe is produced on which a well-defined end face is formed by FIB-milling. In the second step, a nanoscale tip is grown by electron beam-induced deposition of carbon. Next, the carbon tip is coated by a thin layer of chromium to improve adhesion of an aluminum layer evaporated during the final preparation step. Since the length of the tip can be controlled during electron beam deposition, TOA probes can be tailored to provide optimum antenna enhancement for a chosen wavelength [8].

Applications in Nanoscience

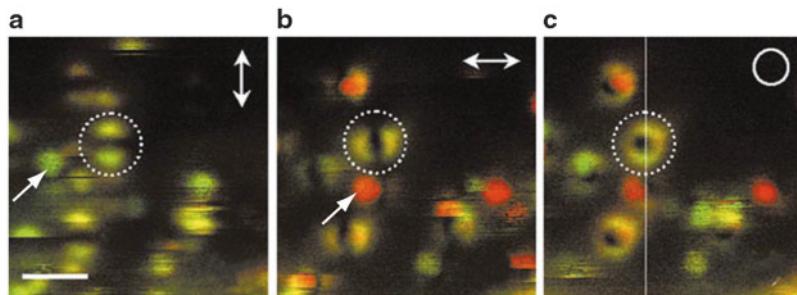
The energy of light quanta – photons – is in the range of electronic and vibrational excitations of materials. These excitations are directly

determined by the chemical and structural composition of matter. Optical spectroscopy, the energy-resolved probing of the material response to light exposure, thus provides a wealth of information on the static and dynamical properties of materials. Combining spectroscopy with near-field microscopy is particularly interesting since spectral information is obtained spatially resolved at the nanoscale. In the following several representative examples covering different material responses including fluorescence, Raman scattering and elastic scattering are briefly illustrated to highlight the capabilities of SNOM techniques.

Fluorescence Microscopy

Fluorescence measurements were among the first applications of SNOM. In these experiments, aperture-based probes were used to probe highly fluorescent dye-molecules on substrates. Aluminum-coated fiber tips were used for excitation while the fluorescence was collected in the far-field by a high-numerical aperture objective. A molecule is excited only if the optical electric field is polarized parallel to its transition dipole moment. The resulting fluorescence patterns rendered by a single molecule with known orientation can thus be used to visualize the local electric field distribution at the probe. Conversely, the molecular orientation can be determined for known field distributions. These experiments showed that the strongest electrical fields do not occur in the center of the aperture, but at the rims of the metal coating. This is the result of local field enhancement at the thin metal rim (see section “[Antenna Probes](#)”). Two lobes with strong fields oriented in axial direction occur located on opposite sides of the aperture in the direction of the polarization of the incident linearly polarized light. Molecules with a transition dipole moment oriented parallel to the tip axis are excited efficiently by these field components as can be seen in Fig. 5 [5]. Rotating the polarization of the incident linearly polarized light is seen to rotate the resulting double lobe pattern that indicates the area with strongest fields.

Fluorescence measurements typically do not require high excitation densities and in many cases, the small light transmission of aperture



Scanning Near-Field Optical Microscopy,
Fig. 5 Series of three successive aperture-SNOM fluorescence images of the same area ($1.2 \times 1.2 \mu\text{m}$) of a sample of dye molecules embedded in a thin transparent polymer film. The excitation polarization, measured in the far-field, was rotated from linear vertical (a) to linear horizontal (b) and then changed to circular polarization (c). Circular features marked by arrows result from molecules with

transition dipole moments oriented parallel to the sample plane. The double lobe structure marked by the dashed circle results from a molecule with perpendicularly oriented transition dipole moment. This molecule senses the strong electrical fields that occur at the rim of the metal aperture at positions determined by the far-field polarization. Scale bar 300 nm (Reprinted with permission from Veerman et al. [5]. Copyright 1999, John Wiley and Sons)

probes represents no major drawback. In fact, fluorescence imaging of single dye molecules with 32 nm spatial resolution has been demonstrated using a microfabricated cantilevered glass tip covered with a 60-nm-thick aluminum film. Due to the thickness of the virtually opaque film, possible contributions from surface plasmons propagating on the outside of the film have been discussed [4]. Examples of fluorescence microscopy measurements based on aperture probes also include, for example, studies of single nuclear pore complexes and the kinetics of protein transport under physiological conditions.

Optical antennas have been used for a variety of samples and materials including photosynthetic proteins, polymers, semiconductor quantum dots, and carbon nano-tubes [11]. The image contrast was based on local field enhancement provided by the tip. The spatial resolution achieved in these experiments was essentially determined by the diameter of the tip and ranged between 10 and 20 nm. Imaging can be combined with local spectroscopy to visualize emission energies on the nanoscale. Single-molecule experiments revealed the field distribution at the tip-antenna in analogy to the discussion made above for aperture probes [14].

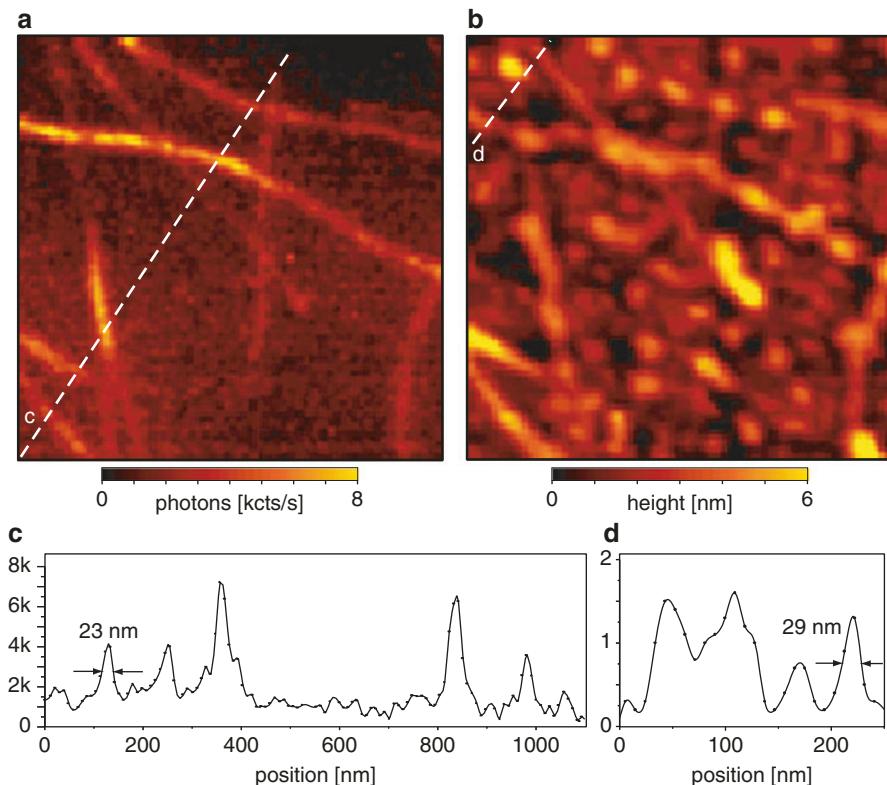
While most of the reported studies were exploiting local signal enhancement, distance-dependent metal-induced quenching of

fluorescence can also be used for high-resolution imaging. This approach provides sub 10 nm spatial resolution and has been applied to single fluorescent organic molecules and inorganic semiconductor nanorods (see e.g., [12]). In these experiments, the spectrally integrated fluorescence signal was demodulated by the tapping-mode frequency of the AFM cantilever after recording photon-arrival times.

Raman Microscopy

Raman scattering probes the unique vibrational spectrum of a sample and directly reflects its chemical composition and molecular structure. A main drawback of Raman scattering is the extremely low scattering cross-section which is typically 10–14 orders of magnitude smaller than the cross-section of fluorescence in the case of organic molecules. Raman measurements thus require higher laser intensities and in many cases the low transmission of aperture probes prohibits their application. The signal enhancement provided by the antenna tip in TENOM is substantial for the detection of nanoscale sample volumina. In the following, a review of selected examples is given to illustrate the possibilities of tip-enhanced Raman scattering (TERS) (see e.g., [11, 15]).

In Fig. 6, simultaneous near-field Raman and topographic imaging of individual single-walled



Scanning Near-Field Optical Microscopy, Fig. 6 Tip-enhanced Raman spectroscopy (TERS) of single-walled carbon nanotubes on glass. Simultaneous near-field Raman image (**a**) and topographic image (**b**). Scan area $1 \times 1 \mu\text{m}^2$. The Raman image is acquired by detecting the intensity of the G' band upon laser excitation at 633 nm. No Raman scattering signal is detected from humidity-related circular features present in the topographic image. (**c**)

Cross-section taken along the *dashed line* in the Raman image indicating a spatial resolution around 25 nm. (**d**) Cross-section taken along the indicated dashed line in the topographic image. Vertical units are photon counts per second for **c** and nanometer for **d** (Reprinted with permission from Hartschuh et al. [16]. Copyright 2003, American Physical Society)

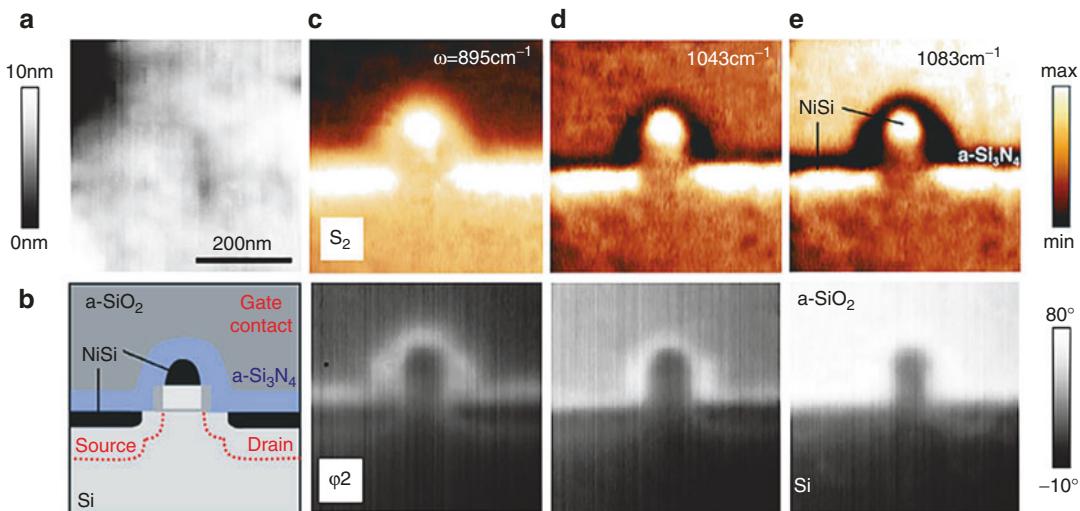
carbon nanotubes is shown. The optical image in (a) reflects the intensity of the G' band, a particular Raman-active vibrational mode of carbon nanotubes. The optical resolution obtained in this experiment was about 25 nm as can be seen from the width of the peaks in the cross-section in Fig. 6c.

The strong fields required for sufficient enhancement of the Raman scattering signal can cause laser-induced decomposition and photochemical reactions in the presence of oxygen. TERS of single electronically resonant molecules has been demonstrated for ultra-high vacuum conditions. A review focusing on single-molecule surface- and tip-enhanced Raman scattering can be found in [15].

Elastic Scattering Microscopy

Elastic scattering SNOM probes the dielectric properties of the sample and has been used from the visible to the microwave regime of the electromagnetic spectrum. Reviews of the fundamentals of the technique and representative applications can be found in [9, 10]. The majority of s-SNOM experiments have been reported for the IR to THz spectral range. Applications include detection of the Mott-transition in nanodomains, mapping of the doping concentration in semiconductors, surface characterization with a sensitivity of a single monolayer, strain-field mapping, and infrared spectroscopy of a single virus.

As an example nanoscale infrared spectroscopic near-field mapping of single nano-



Scanning Near-Field Optical Microscopy.
Fig. 7 Material-specific mapping of transistor components using s-SNOM: Cross-sectional images of a single transistor fabricated at the 65 nm technology node. (a) Topography. (b) Sketch of the transistor with materials indicated. (c–e) Near-field amplitude and phase images

transistors is shown in Fig. 7. A cantilevered metallized Si-tip operating in tapping-mode with an oscillation frequency of 300 kHz and an amplitude of about 60 nm was used [17]. The data clearly demonstrates the potential of s-SNOM for infrared spectroscopic recognition of materials within individual semiconductor nanodevices.

Based on the antenna approach, s-SNOM typically provides 10–20 nm spatial resolution determined by the diameter of the tip-apex. In most of the s-SNOM experiments to date, monochromatic laser sources were used. Since only the optical response at this frequency is determined, the acquisition of scattering spectra or spectrally resolved images can only be obtained sequentially with a series of image scans at different laser frequencies. New developments exploiting broadband NIR laser sources aim at overcoming this limitation, and recently obtained a spectral bandwidth exceeding 400 cm^{-1} .

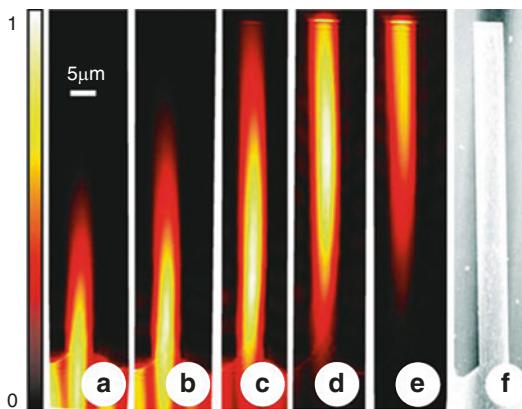
Plasmonics and Photonic Nanostructures

SNOM plays a vital role in the field of plasmonics which deals with the study of optical phenomena related to the electromagnetic response of metals [18]. Near-field optical probes are particularly

recorded at three different laser frequencies. Amorphous SiO_2 and Si_3N_4 render reversed optical contrast and are clearly distinguished. A spatial resolution better than 20 nm has been achieved (Reprinted with permission from Huber et al. [17]. Copyright 2010, IOP)

important for two reasons. First, they provide a means to locally excite propagating surface plasmon polaritons (SPPs) in metal films, a process that is not possible in case of propagating light waves because of momentum (k -vector) mismatch. The broad k -spectrum associated with the near-field of the probe contains sufficient bandwidth for efficient SPP-excitation (see Fig. 1). Second, near-field probes can be used simultaneously to convert SPPs back into propagating waves, thereby probing the local distribution of electromagnetic fields in the vicinity of metallic nanostructures. As an example, the near-field associated with SPPs has been visualized along gold nanowires using an aperture probe in collection mode (see Fig. 2a, [4]). This approach is also termed photon scanning tunneling microscopy (PSTM) to illustrate the analogy between evanescent electromagnetic waves and the corresponding exponentially decaying electron wavefunctions within the tunnel barrier of an STM. PSTM has been widely used to spatially resolve light wave propagation also in dielectric photonic nanostructures [4].

Besides the visualization of static field distributions, optical spectroscopy also allows for the



Scanning Near-Field Optical Microscopy,

Fig. 8 Phase-sensitive and ultrafast near-field microscopy of a surface plasmon polariton (*SPP*) waveguide. The local electric field is collected by an aperture-probe and detected interferometrically in a Mach-Zehnder-type configuration. (a–e) Normalized amplitude information of the SPP wavepacket E-field. Succeeding frames are new scans of the probe. In between the frames the delay line is lengthened to 14.4 μm . Therefore, the time between two frames is 48 fs. The scan frame is $15 \times 110 \mu\text{m}^2$, scan lines run from *top* to *bottom*. (f) Topography of the SPP waveguide obtained by shear-force feedback (Reprinted with permission from Sandtke et al. [19]. Copyright 2008, American Institute of Physics)

study of their temporal evolution and the propagation of pulses. Figure 8 illustrates ultrafast and phase-sensitive imaging of the plasmon propagation in a metallic waveguide by PSTM [19]. In this case the near-field microscope uses an aperture-probe in collection mode and incorporates a Mach-Zehnder-type interferometer enabling heterodyne time-resolved detection.

Scattering-SNOM with antenna tips has been used extensively to study localized surface plasmon polaritons (LSPP) in different metal nanostructures. By varying the laser excitation frequency near-field optical imaging allowed for distinguishing higher order plasmonic resonances [20].

Perspectives

During the last 25 years SNOM has demonstrated its capabilities for sub-wavelength optical

imaging and spectroscopy of surfaces and sub-surface features. The strength of SNOM results from its enormous flexibility with respect to sample types as well as measurement configurations and in particular, from its combination with a broad range of spectroscopic techniques. Ongoing developments aim at increasing antenna efficiencies and new aperture-type schemes [18]. In addition, the combination of nano-optical approaches and ultrafast laser technique is explored to achieve enhanced light localization and the control of optical near-fields on the time scale of few optical cycles.

Cross-References

- [Atomic Force Microscopy](#)
- [Confocal Laser Scanning Microscopy](#)
- [Light Localization for Nano-optical Devices](#)
- [Nanostructures for Photonics](#)
- [Scanning Tunneling Microscopy](#)

References

1. Novotny, L.: The history of near-field optics. In: Wolf, E. (ed.) *Progress in Optics*, vol. 50, pp. 137–184. Elsevier, Amsterdam (2007)
2. Novotny, L., Hecht, B.: *Principles of Nano-optics*. Cambridge University Press, Cambridge (2006)
3. Kawata, S., Shalaev, V.M. (eds.): *Advances in Nano-optics and Nano-photonics Tip Enhancement*. Elsevier, Amsterdam (2007)
4. Kawata, S., Shalaev, V.M. (eds.): *Handbook of Microscopy for Nanotechnology*. Kluwer, Dordrecht (2006)
5. Veerman, J.A., Garcia-Parajo, M.F., Kuipers, L., van Hulst, N.F.: Single molecule mapping of the optical field distribution of probes for near-field microscopy. *J. Microsc.* **194**, 477–482 (1999)
6. Mihalcea, C., Scholz, W., Werner, S., Münster, S., Oesterschulze, E., Kassing, R.: Multipurpose sensor tips for scanning near-field microscopy. *Appl. Phys. Lett.* **68**, 3531–3533 (1996)
7. Hecht, B., Sick, B., Wild, U.P., Deckert, V., Zenobi, R., Martin, O.J.F., Pohl, D.E.: Scanning near-field optical microscopy with aperture probes: fundamentals and applications. *J. Chem. Phys.* **112**, 7761–7774 (2000)
8. Bharadwaj, P., Deutsch, B., Novotny, L.: Optical antennas. *Adv. Opt. Photon.* **1**, 438–483 (2009)

9. Keilmann, F., Hillenbrand, R.: Near-field microscopy by elastic light scattering from a tip. *Philos. Trans. R. Soc. Lond. A* **362**, 787–805 (2004)
10. Bründermann, E., Havenith, M.: SNIM: scanning near-field infrared microscopy. *Annu. Rep. Prog. Chem., Sect. C: Phys. Chem.* **104**, 235–255 (2008)
11. Hartschuh, A.: Tip-enhanced near-field optical microscopy. *Angew. Chem. Int. Ed.* **47**, 8178–8198 (2008)
12. Ma, Z., Gerton, J.M., Wade, L.A., Quake, S.R.: Fluorescence near-field microscopy of DNA at sub-10 nm resolution. *Phys. Rev. Lett.* **97**, 260801–260804 (2006)
13. Taminiau, T.H., Stefani, F.D., Segerink, F.B., van Hulst, N.F.: Optical antennas direct single-molecule emission. *Nat. Photonics* **2**, 234–237 (2008)
14. Frey, H.G., Witt, S., Felderer, K., Guckenberger, R.: High resolution imaging of single fluorescent molecules with the optical near field of a metal tip. *Phys. Rev. Lett.* **93**, 200801–200804 (2004)
15. Pettinger, B.: Single-molecule surface-and tip-enhanced Raman spectroscopy. *Mol. Phys.* **108**, 2039–2059 (2010)
16. Hartschuh, A., Sánchez, E.J., Xie, X.S., Novotny, L.: High-resolution nearfield Raman microscopy of single-walled carbon nanotubes. *Phys. Rev. Lett.* **90**, 095503–4 (2003)
17. Huber, A.J., Wittenborn, J., Hillenbrand, R.: Infrared spectroscopic near-field mapping of single nanotransistors. *Nanotechnology* **21**, 235702–6 (2010)
18. Schuller, J.A., Barnard, E.S., Cai, W., Jun, Y.C., White, S.W., Brongersma, M.L.: Plasmonics for extreme light concentration and manipulation. *Nat. Mater.* **9**, 193–204 (2010)
19. Sandtke, M., Engelen, R.J.P., Schoenmaker, H., Attema, I., Dekker, H., Cerjak, I., Korterik, J.P., Segerink, F.B., Kuipers, L.: Novel instrument for surface plasmon polariton tracking in space and time. *Rev. Sci. Instrum.* **79**, 013704–10 (2008)
20. Dorfmüller, J., Vogelgesang, R., Khunzin, W., Rockstuhl, C., Etrich, C., Kern, K.: Plasmonic nanowire antennas: experiment, simulation, and theory. *Nano Lett.* **10**(9), 3596–3603 (2010)

Scanning Probe Microscopy

► [Atomic Force Microscopy](#)

Scanning Surface Potential Microscopy

► [Kelvin Probe Force Microscopy](#)

Scanning Thermal Microscopy

Li Shi

Department of Mechanical Engineering,
The University of Texas at Austin, Austin,
TX, USA

Synonyms

[Microthermal analysis](#); [Scanning thermal profiler](#)

Definition

Scanning Thermal Microscopy (SThM) is a class of experimental methods for high spatial resolution mapping of the surface temperature distribution of an operating device or the thermal property variation of a structure with the use of a sensor fabricated on a scanning probe.

Thermal Probes

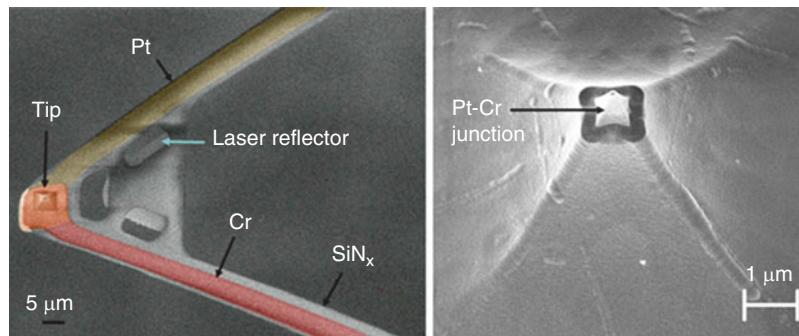
In 1986, Williams and Wickramasinghe [1] pioneered a so-called scanning thermal profiler technique based on a thermocouple sensor fabricated at the end of a probe tip for Scanning Tunneling Microscopy (STM). The thermocouple sensor consisted of two dissimilar conductors that made a junction at the end of the STM tip. An insulator separated the two conductors in all areas remote from the tip. The size of the thermocouple junction could be made as small as 100 nm. When a temperature difference (ΔT) exists between the thermocouple junction and the ends of the two lead wires, thermal diffusion of electrons through the two dissimilar wires results in a thermoelectric voltage between the two lead wires. The magnitude of the thermovoltage measured with the use of a voltmeter is

$$V = (S_1 - S_2)\Delta T \quad (1)$$

where S_1 and S_2 are the Seebeck coefficient of the two thermocouple wires.

Scanning Thermal Microscopy,

Fig. 1 Scanning electron micrographs of a SiN_x AFM cantilever probe (**a**) with a sub-micron Pt-Cr thermocouple junction (**b**) formed at the apex of the SiO_2 tip



The main purpose of this first scanning thermal profiler was not for mapping temperature distribution of a surface but to regulate the tip-sample distance using the relationship between the thermocouple tip temperature and the distance between the tip and the sample. The work of Williams and Wickramasinghe stimulated intense efforts to develop a scanning probe for high spatial resolution mapping of the temperature distribution or thermal properties on a surface. In 1993, Majumdar and co-workers [2] introduced a wire thermocouple probe that could be used in an atomic force microscope (AFM) for simultaneous mapping of topography and temperature. Since then, different designs of scanning thermal probes have been fabricated. A common feature of these thermal probes is a thermal sensor fabricated at the end of an AFM or STM tip. The thermal sensor can be a thermocouple, a resistance thermometer, a Schottky diode, or a fluorescence particle [3–5].

As discussed above, a thermocouple measures the temperature difference between the junction of its two constituent metals and the other ends of the two metal wires. The current–voltage (I–V) characteristics of a Schottky junction depends on the temperature, and can thus serve as a local temperature sensor. A resistance thermometer is usually made of a metal or degenerately doped semiconductor with a relatively large and constant temperature coefficient of resistance. Because the resistance of a nanoscale resistor can be too small for sensitive electronic detection, resistance thermometer cannot be miniaturized as readily as a thermocouple sensor. Another approach is to use the emission band of some fluorescence particles, which shifts with increasing temperature. This

feature has been used for making a SThM probe with a fluorescence particle located at the end of a scanning probe tip [5]. Besides these temperature-sensing techniques, the thermally induced bending of a biomaterial AFM cantilever has been used for temperature measurements [4]. Moreover, the thermal expansion of a sample has been measured with a regular AFM tip, and used to infer the sample temperature [4].

Different methods have been reported for fabricating different SThM probes. Some of the methods use sequential (or one at a time) fabrication methods [4], whereas batch fabrication processes have been developed for wafer scale fabrication of the SThM probes [6]. While the probe batch fabrication processes are somewhat similar to those that have been developed for the manufacturing of ultra large scale integrated (USLI) devices, they often involve additional etching steps to make free standing cantilever probes, as well as novel processes to form a thermal sensor such as a thermocouple at the end of the cantilever probe tip. Figure 1 shows a batch-fabricated thermocouple SThM cantilever probe that consists of a 0.5 μm thick low stress silicon nitride (SiN_x) cantilever, a 8 μm tall silicon dioxide (SiO_2) pyramid tip with a ~ 20 nm tip radius, and a sub-micron Pt-Cr thermocouple junction formed at the apex of the SiO_2 tip [6]. Several other unique fabrication processes have been developed for wafer scale fabrication of a thermocouple junction, Schottky diode, Pt-C or doped silicon resistance thermometer [4].

Special probe holders with necessary electrical contacts or optical access can allow for the use of the SThM sensor probes in a commercial AFM or STM. STM requires the sample surface to be

conducting because the tip-sample gap is controlled based on the tunneling current. This issue has limited the use of SThM probes in STM. In comparison, AFM with a cantilever SThM probe can allow simultaneous topography and thermal mapping of both conducting and non-conducting samples. In a typical contact-mode AFM, the tip-sample spacing is regulated by the force acting on the probe tip. The tip-sample interaction force can be obtained from the AFM cantilever bending that is measured with the use of several methods. One of these methods is to measure the laser beam reflected by the cantilever with the use of a position-sensitive detector. Another method is to employ a built-in piezoresistive sensor in the cantilever to measure the cantilever bending. Although the optical detection method provides superior sensitivity, the laser heating of the thermal sensor at the end of the cantilever needs to be minimized.

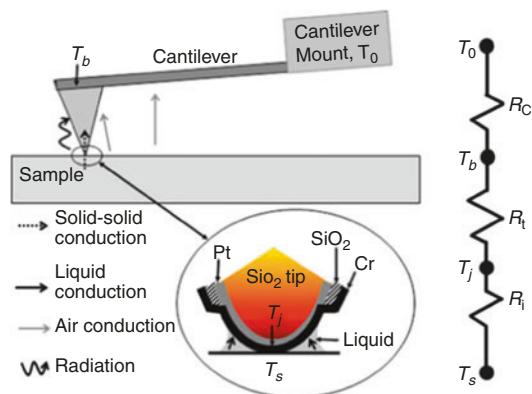
Theoretical Analysis of Heat Transfer Mechanisms

For many of the various SThM sensor probes, the sensitivity and spatial resolution of the thermal imaging technique depend on the mechanisms of heat transfer between the thermal probe and the sample. For SThM measurements conducted in air, heat is transferred between the probe and the sample via conduction through the solid-solid contact, a liquid meniscus at the tip-sample contact, and the air gap between the probe, and via radiation, as illustrated in Fig. 2 heat transfer through the air gap and via radiation is not localized at the tip-sample contact, and can deteriorate the spatial resolution of the SThM probes.

For a circular contact area with radius b , the spreading thermal resistance of the sample is obtained as [7]

$$R_s = \frac{1}{4\kappa_s b} \quad (2)$$

where κ_s is the thermal conductivity of the sample. For $\kappa_s \approx 5 \text{ W/m-K}$, $b \approx 30 \text{ nm}$, $R_s \approx 1.7 \times 10^6 \text{ K/W}$.



Scanning Thermal Microscopy, Fig. 2 Schematic diagram showing the heat transfer mechanisms between the thermal probe and the sample as well as a thermal resistance circuit where the heat transfer through the air gap is ignored (Reproduced from [6] with permission by ASME)

The thermal interface resistance through the solid-solid contact is given as [8]

$$R_{i,s} \approx \frac{4}{\alpha C_s v_s \pi b^2} = \frac{4K}{3\alpha \kappa_s \pi b} \quad (3)$$

where C_s and v_s are the specific heat and phonon group velocity of the sample, α is the phonon transmission coefficient from the sample into the tip, $K \equiv l_s/b$ is the Knudsen number, and l_s is the phonon mean free path in the sample. The second equation is obtained with the use of the kinetic theory, $\kappa_s = C_s v_s l_s / 3$ [8]. For $\alpha = 0.3$, $\kappa_s = 5 \text{ W/m-K}$, $b = 30 \text{ nm}$, and $l_s = 5 \text{ nm}$, $R_{i,s} = 1.5 \times 10^6 \text{ K/W}$.

A liquid meniscus often forms at the tip-sample junction when the tip is scanned on the sample in air. The interface thermal resistance through the liquid meniscus ($R_{i,l}$) at the tip-sample junction depends on the humidity and surface properties. The value of $R_{i,l}$ has been estimated by Majumdar to be on the order of 10^5 K/W [4], which is lower than the solid-solid thermal interface resistance. The total interface tip-sample thermal resistance is

$$R_i = \left(R_{i,s}^{-1} + R_{i,l}^{-1} \right)^{-1} \quad (4)$$

The spreading thermal resistance of the conical tip can be estimated as

$$R_t \approx \frac{1}{\pi \kappa_t b \tan \theta} \quad (5)$$

where κ_t is the thermal conductivity of the tip and θ is the half angle of the conical tip. For $\kappa_t \approx 5 \text{ W/m-K}$, $b = 30 \text{ nm}$, $\theta = \pi/8$, $R_t = 5 \times 10^6 \text{ K/W}$. Hence, if the thermocouple junction can be made as small as the contact size, the high spreading thermal resistance of the tip can be utilized to thermally isolate the junction from the cantilever.

An approximation expression for the tip-sample air thermal conductance per unit area is given here as

$$g_a \approx \frac{\kappa_a}{z + l_a} \quad (6)$$

where κ_a and l_a are the thermal conductivity and mean free path of air molecules, and z is the air gap. In the ballistic limit of $z \ll l_a$, this expression is reduced to $g_a = \kappa_a/l_a$. Based on the kinetic theory, $\kappa_a = C_a v_a l_a/3$, where C_a and v_a are the specific heat and velocity of air molecules, so that $g_a = C_a v_a/3$ for the ballistic limit. This result is close to the ballistic thermal conductance per unit area [8], $C_a V_a/4$, for heat transfer of gas molecules between two close parallel plates when the thermal accommodation coefficient is unity, corresponding to the case that the scattered molecules take the temperature of the surface. In the diffusive limit of $z \gg l_a$, $g_a = \kappa_a/z$ is the diffusive thermal conductance per unit area of the air gap. The total tip-sample thermal conductance through the air gap can be obtained as

$$\begin{aligned} G_{g,t-s} &= \int_b^{b+H \tan \theta} \frac{\kappa_a}{l_a + (r - b)/\tan \theta} 2\pi r dr \\ &= 2\pi \kappa_a H \tan^2 \theta \left[1 + \frac{b - l_a \tan \theta}{H \tan \theta} \ln \frac{l_a + H}{H} \right] \end{aligned} \quad (7)$$

where H is the tip height. For $H = 8 \mu\text{m}$ and $\theta = \pi/8$, $G_{g,t-s} \approx 2.2 \times 10^{-7} \text{ W/K}$, corresponding to a resistance $R_{g,t-s} \equiv 1/G_{g,t-s} = 4.4 \times 10^6 \text{ K/W}$ [6].

Far-field radiation conductance between the tip and the sample is approximated as that of a two-surface enclosure as

$$G_{\text{rad},t-s} \approx \frac{4A_t T^3}{(1 - \varepsilon_t)/\varepsilon_t + 1/F_{t-s} + A_t(1 - \varepsilon_s)/A_s \varepsilon_s} \quad (8)$$

where σ is the Stefan–Boltzmann constant, A_t and A_s are the surface areas of the tip and sample, ε_t and ε_s are the surface emissivity of the tip and sample, and F_{t-s} and T are the view factor and the average temperature between the tip and sample. When the temperature is close to 300 K, $G_{\text{rad},t-s}$ is on the order of $1 \times 10^{-10} \text{ W/K}$. Both the far-field and near-field radiation heat transfer between the tip and the sample can be ignored compared to those via the air gap, solid–solid interface, and liquid meniscus [4], unless for a very high tip or sample temperature.

Because of heating by the sample surface, the temperature of the air molecules surrounding the cantilever, T_g , can be different from the temperature at the cantilever mount, T_0 , which is usually close to the room temperature. In this case, heat transfer from the tip into the cantilever can be obtained as the fin heat transfer rate expressed as [7]

$$q_{t-c} = M \frac{\cosh mL - \theta_0/\theta_b}{\sinh mL} \quad (9)$$

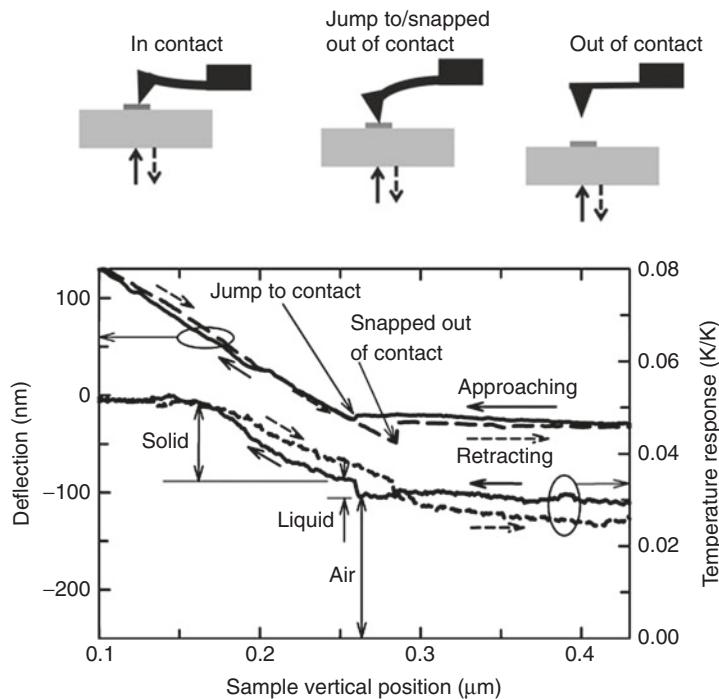
where $M = (hP\kappa_c A_c)^{1/2}\theta_b$, $m = (hP/\kappa_c A_c)^{1/2}$, h is the heat transfer coefficient between the cantilever and the surrounding gas, κ_c and A_c are the thermal conductivity and cross-section of the cantilever, P is the perimeter of the cantilever cross-section, $\theta_0 = T_0 - T_g$, $\theta_b = T_b - T_g$, and T_b is the temperature at the joint between the tip and the cantilever. When only a small sample area is at a temperature different from the room temperature so that $T_g \approx T_0$, the thermal resistance of the cantilever can be defined as

$$R_c = \frac{T_b - T_0}{q_{t-c}} = \frac{\tanh mL}{\sqrt{hP\kappa_c A_c}} \quad (10)$$

For $\kappa_c \approx 5 \text{ W/m-K}$, and $0.5 \mu\text{m}$ by $10 \mu\text{m}$ cantilever cross-section, $L = 150 \mu\text{m}$, $h \approx k_a(\pi/A_c)^{1/2}$, $R_c = 3 \times 10^5 \text{ K/W}$.

Scanning Thermal Microscopy,

Fig. 3 Cantilever deflection and temperature response of a thermocouple probe as a function of the sample vertical position when a 350-nm wide heater line sample was raised toward and then retracted from the tip (Reproduced from [6] with permission by ASME)



On the other hand, when a large area of the sample is at a temperature higher than T_0 , the air molecules surrounding the cantilever can be heated by the sample surface to a temperature T_g that is considerably higher than T_0 . In this case, the cantilever is heated by the surrounding air, resulting in an additional heat transfer path from the sample to the probe.

Experimental Characterization of Heat Transfer Mechanisms

Experimentally, the heat transfer path through the air gap between the sample and the cantilever was found to be significant when the size of the heated zone on the sample surface was not small [6]. In the experiment [6], a 350-nm wide metal line was joule heated to 5.3 K above room temperature. The cantilever deflection and temperature rise of the thermocouple junction were recorded simultaneously when the sample was approached and then retracted from a thermocouple SThM probe tip. When the sample approached the tip, the cantilever deflection signal remained

approximately constant before the sample contacted the tip, as shown in the deflection curve in Fig. 3. In this region, the junction temperature rise was caused by air conduction between the probe and the sample. As the tip-sample distance was reduced, the junction temperature rise due to air conduction increased slowly. Before the sample made solid-solid contact to the tip, the adsorbed liquid layers on the tip and the sample bridged each other. Initially, this liquid bridge pulled the tip down by a van der Waals force, as being seen in the dip labeled as “jump to contact” in the deflection curve. Coincidentally, there was a small jump in the junction temperature due to heat conduction through the liquid bridge. As the sample was raised further, both the solid-solid contact force and the junction temperature increased gradually, until the cantilever was deflected for more than 100 nm. After this point, the junction temperature remained almost constant as the contact force increased.

As the sample was retracted from the tip, the junction temperature remained almost constant until at a cantilever deflection of 100 nm, after which the junction temperature rise decreased

roughly linearly but at a smaller slope than that found in the approaching cycle. As the sample was lowered further, the tip was pulled down together with the sample by surface tension of the liquid bridge until after a certain point, the restoring spring force of the cantilever exceeded the surface tension, and the tip “snapped out of contact” with the sample. Associated with the breaking of the liquid bridge, there was a small drop in the junction temperature.

This experiment shows several mechanisms. First, before the tip contacted the sample, air conduction contributed to a junction temperature rise up to 0.03 K per K sample temperature rise, which was about 60 % of the maximum junction temperature rise at the maximum contact force of the experiment. Second, conduction through a liquid meniscus was responsible for the sudden jump and drop in junction temperature when the tip “jumped to contact” to and “snapped out of contact” from the sample, respectively. Third, solid–solid conduction resulted in the almost linear relationship between the junction temperature rise and the contact force. This is a well understood feature for macroscopic solid–solid contacts. When the contact force was increased further, the junction temperature approached a constant likely because the contact size approached the maximum possible value limited by the diameter of an asperity located at the tip-sample interface.

The point contact experiment was repeated for a 5.8 μm wide and 2,000 μm long heater line on an oxidized silicon wafer. While the increase of the normalized junction temperature due to solid and liquid conduction was similar in magnitude to those observed in the narrower line, the normalized junction temperature rise due to air conduction was one order of magnitude higher than the corresponding one in the narrower line. In addition, the temperature rise of the thermocouple junction was measured when the tip was in contact with three heater lines of different line widths. It was found that the temperature rise at the thermocouple junction of the tip was about 53 %, 46 %, and 5 % of that of a 50, 3, and 0.3 μm wide line, respectively, showing a trend of increasing temperature rise in the SThM probe with increasing

heater line width. These results suggest that the cantilever was heated more by air conduction between the tip and the larger hot area of the wider heater lines. This trend indicates that air conduction plays an important role in probe–sample heat transfer, especially when the hot area on the sample surface is large in comparison to the tip size.

The unwanted heat transfer through the air gap can be eliminated by conducting SThM measurements in vacuum. In this case, the tip-sample air conductance $G_{g, t-s} = 0$, and the thermal resistance of the cantilever is the conduction thermal resistance given as

$$R_c = \frac{L}{\kappa_c A_c} \quad (11)$$

For a cantilever length $L = 150 \mu\text{m}$, thermal conductivity $\kappa_c = 5 \text{ W/m-K}$, and cross-section $A_c = 0.5 \mu\text{m} \times 10 \mu\text{m}$, $R_c = 6 \times 10^6 \text{ W/K}$ based on the above equation. In addition, the heat transfer from the sample to the tip in vacuum can be described with the use of the thermal resistance circuit shown in Fig. 2.

However, vacuum operation of the SThM probe is considerably much more challenging than operation in open air environment. Moreover, the liquid meniscus at the tip-sample junction is also eliminated by the vacuum environment. Although this can further enhance the spatial resolution of the thermal imaging technique, the tip-sample thermal interface conductance could be reduced considerably. Consequently, the thermocouple junction temperature at the tip end could become rather different from the sample temperature, resulting in low sensitivity and large uncertainty in the measured temperature.

An alternative method to eliminate the unwanted air conductance while keeping the heat transfer via the liquid meniscus is based on a dual scan measurement approach [9]. In this approach, after a line scan of the SThM probe on the sample surface in the contact mode, the probe is scanned on the same line in the lift mode with a small fixed tip-sample distance between the tip and the sample. The difference between the two thermal sensor signals obtained in the contact mode and the lift

mode is attributed to the local heat transfer through the solid–solid contact and the liquid meniscus at the tip-sample junction, and can be used to obtain the local sample surface temperature at the contact point, as discussed below.

Surface Temperature Mapping

When the heat transfer through the probe-sample gap is eliminated by the dual scan technique, the difference in the measured thermocouple junction temperature values in the contact mode and the lift mode is still different from that for vacuum measurement given by the thermal resistance circuit of Fig. 2 as

$$T_j = (T_s - T_0) \left[1 + \frac{R_i}{R_c + R_t} \right]^{-1} + T_0 \quad (12)$$

where T_s is the local sample temperature at the contact point, T_0 is the ambient temperature.

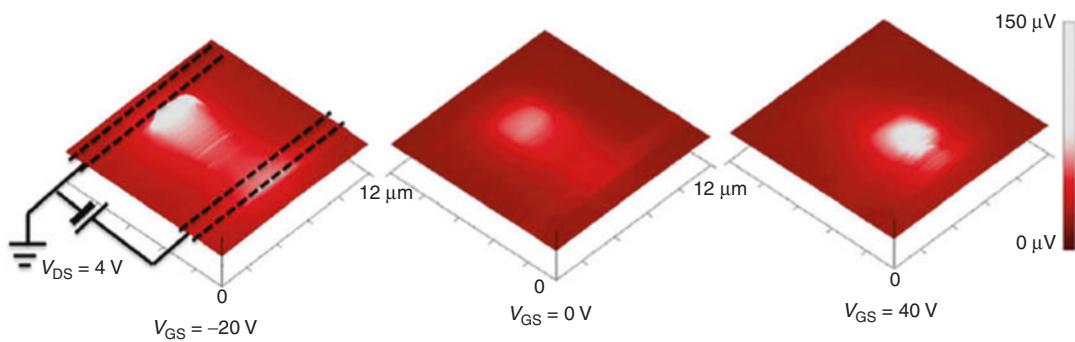
It is important to increase R_t and R_c relative to R_i and R_s , so as to maximize ΔT_j and minimize the sample temperature change due to the contact by the tip. For this reason, it is desirable to use low-thermal conductivity dielectrics such as SiO₂ and SiN_x to fabricate the pyramid tip and the cantilever, and low-thermal conductivity conductors such as Cr and Pt to fabricate the thermocouple sensor. The geometry of the sensor, tip, and cantilever also needs to be optimized to minimize heat loss.

Even for state-of-the-art thermocouple SThM probes, R_t and R_c are still not much larger than R_i . Hence, the ΔT_j obtained from the differential dual scan measurement needs to be calibrated in a separate experiment, where the sample is heated with an external heater and the sample surface temperature is determined using a different method, such as a resistance thermometer fabricated on the sample surface [10].

SThM methods operated in the contact mode or the dual scan mode have been employed to measure the surface temperature distribution of silicon, carbon nanotube, and graphene nanoelectronic devices, with a spatial resolution better than 100 nm [10, 11]. The very high electric

field in electrically biased nanoelectronic devices can result in non-equilibrium transport of electrons, optical phonons, and acoustic phonons [12]. Energetic hot electrons accelerated in the high field lose their energy to lattice vibration via emission of high frequency optical phonons. The optical phonons are further scattered with acoustic phonons that make a larger contribution to the lattice thermal conductivity because of a much higher group velocity than the optical phonons. The scattering time between optical phonons and acoustic phonons can be much longer than that between hot electrons and optical phonons in nanoelectronic devices made of silicon and carbon nanotubes. Consequently, energy transfer from optical phonons and acoustic phonons becomes a bottle neck in the heat dissipation process. This bottle neck can result in different temperatures for acoustic phonons and optical phonons. The local temperatures of different energy carriers in operating nanoelectronic devices can be measured with different methods. For example, infrared spectroscopy has been used to measure the electronic temperature of electrically biased graphene devices by fitting the infrared emission from the hot electrons in the device with the Planck distribution [13]. The intensity ratio between the anti-Stokes peak to the Stokes peak in the Raman spectrum can be used to probe the local temperature of the zone center or zone boundary optical phonons that are active in the Raman scattering processes [14]. Complementary to these optical thermometry methods, the thermocouple SThM probe can be used to map the low frequency phonon temperature distribution because the phonon transmission coefficient at the tip-sample interface increases with decreasing phonon frequency. This capability has been employed to observe bias-dependant and asymmetric acoustic phonon temperature distribution in electrically biased graphene with a superior spatial resolution of about 100 nm [10], as illustrated in Fig. 4.

Besides the dual scan method, an active SThM method has been reported by Nakabeppu and Suzuki [15] to obtain quantitative temperature map of a sample surface in vacuum. Their probe for active SThM consists of two thin film



Scanning Thermal Microscopy, Fig. 4 Measured SThM thermovoltage maps of an electrically biased graphene device for a constant drain-source bias

($V_{DS} = 4$ V) and different gate-source voltage of -20 V, 0 V, and 40 V, respectively. The scan size is $12 \mu\text{m} \times 12 \mu\text{m}$

thermocouples (TCs) and a micro-heater fabricated on a cantilever. When the cantilever was scanned on a heated sample, heat flow from the sample into the cantilever resulted in a temperature drop (ΔT) along the cantilever, which was measured using the differential thermocouple. The micro-heater power was adjusted by using a feedback loop until the differential thermocouple reads $\Delta T = 0$. Under this condition, the heat flow from the sample to the cantilever was zero so that the measured tip temperature is the same as the sample surface temperature. Alternatively, the active SThM method can also be implemented using the dual scan operation in air [9]. In this implementation, the SThM probe is heated to different temperatures and is scanned on a hot sample at each heating rate for the probe. When the measured sensor temperature of the heated SThM probe is the same during the contact-mode and lift-mode scanning of the probe, the sensor temperature is taken as the same as the local sample temperature, and the heat flow between the tip and the sample is nullified.

It is worth emphasizing that the key issue in SThM measurements is that the thermal sensor temperature can be quite different from the sample temperature. This issue was addressed in the aforementioned measurements via either a detailed calibration or actively heating the sensor until the sensor temperature was the same as the sample temperature. While four-point measurement methods with minimum current leakage

into the voltage measurement devices are commonly used to address problems caused by contact electrical resistance in electrical measurements, a thermal analogue of this approach has not been effective for addressing the problem caused by thermal contact resistance in SThM measurements, because it is difficult to eliminate heat loss into a temperature measurement device, or the thermal probe in SThM.

Thermal Property Mapping

Scanning thermal probes consisting of a resistance heater and thermometer at the tip end have been used to measure the local thermal property near the surface in a solid sample. For this measurement, the temperature of the heated probe tip decreases as the tip touches the sample. A larger drop in the tip temperature is expected when the tip touches a higher thermal conductivity region of the sample because of a smaller spreading thermal resistance in the sample. Alternatively, the probe heating rate can be feedback-controlled to maintain a constant tip temperature during tip scanning. In this case, a higher heating power is needed to achieve a constant tip temperature when the tip touches a higher thermal conductivity region. Although thermal conductivity contrast has been observed under dc operations, accurate thermal conductivity measurement remains a challenge especially for low-thermal conductivity samples [4]. This is in part because a significant

fraction of the power may be dissipated through the cantilever instead of into the sample. The heat loss through the cantilever can be reduced and the spatial resolution can be improved by using microfabricated probes with a low-thermal conductivity cantilever and a sharp pyramid Si resistance heater and thermometer tip. However, with a decreased contact radius b , the tip-sample thermal interface resistance increases according to $1/b^2$, and can dominate the spreading thermal resistance of the sample that scales as $1/b$. The large interface thermal resistance can make it difficult to quantify the sample thermal property when the tip size is small.

The resistance thermometer probe can also be operated in the ac mode, as shown by Pollock, Hammiche, and coworkers [3]. The sensor probe used in their measurements was 5- μm -diameter Pt resistance wire bent to form a tip at the bent. An ac electrical heating current was used to modulate the probe temperature by 5 °C at 10 kHz around a dc temperature of 40 °C, the measured amplitude and the phase shift in the ac voltage drop across their Pt wire thermometer probe depended on the thermal properties of the sample. One advantage of measuring the phase lag is that it is independent of the temperature dependence of the electrical resistivity of Pt wire and the power input to the probe, because the phase lag is given as [4]

$$\tan \phi = -\frac{G_{\text{im}}}{G_r} \quad (13)$$

where G_r and G_{im} are the real and imaginary parts of the complex AC thermal conductance, which can be expressed as

$$G_r \approx 2\pi\kappa_s b \left(1 + b \sqrt{\frac{\omega}{2\alpha_s}} \right) + A_c \kappa_c \sqrt{\frac{\omega}{2\alpha_c}} \quad (14)$$

$$G_{\text{im}} \approx \omega mC + 2\pi\kappa_s b^2 \sqrt{\frac{\omega}{2\alpha_c}}$$

where b is the contact radius, κ_s and α_s are the thermal conductivity and thermal diffusivity of the sample, κ_c and α_c are the thermal conductivity and thermal diffusivity of the cantilever probe,

and the mC product is the thermal mass of the temperature sensor. For simplicity, the thermal interface resistance is ignored in this analysis [4]. For the phase lag to be sensitive to the thermal properties of the sample, the thermal conductance of the cantilever probe cannot be much larger than that of the sample. In addition, it should be noted that both the amplitude and the phase of the signal are influenced by the penetration depth of $\sqrt{2\alpha_s/\omega} = \delta$. Hence, varying the frequency can allow for depth profiling and sub-surface imaging.

Pollock, Hammiche, and coworkers [3] have also used their Pt wire resistance heater and thermometer probe to perform localized ac calorimetry [16]. Here they ramped the temperature of both the sample and the probe at about 15 °C/min while adding an ac temperature modulation of 1 °C at 10 kHz by the probe. They found that any phase transition in the sample gave rise to a change in the phase signal in ac mode. This is observed more clearly in the first derivative of the phase as a function of temperature. By scanning the sample under this mode, calorimetric analysis can be performed locally.

Summary and Future Directions

A number of SThM probes have been designed and fabricated. Different operating methods including the dual scan and the zero heat flux techniques have been employed for quantitative mapping of the surface temperature distribution of operating devices. Both dc and ac operations of a resistance heater and thermometer probe have allowed the mapping of local thermal property variation of nanostructured materials. Further enhancement of the spatial resolution and thermal measurement accuracy of the SThM methods can be made by continuous miniaturization of the sensor size at the probe tip as well as improved thermal isolation of the sensor. Detecting the phase lag in the ac SThM signal can potentially allow for locating sub-surface defects that generate localized heating in operating electronic devices, as suggested in [17].

The SThM probes based on a thermal-to-electrical signal transduction method can be complemented with non-contact optical methods for mapping the local temperature distribution of different phonon populations or different thermal properties. Although the spatial resolution of conventional far-field optical imaging methods is limited by diffraction to be on the order of the wavelength, the diffraction barrier has been broken by near-field scanning optical microscopy (NSOM) methods [18] as well as several far-field optical nanoscopy methods [19]. Near-field infrared evanescent wave emitted from a sample surface has been collected using a solid immersion lens [11] and/or scattered into far-field signal by an apertureless metal tip [20]. These techniques have the potential for profiling the surface temperature distribution of devices with a spatial resolution of a fraction of the IR wavelength. Similarly, the spatial resolution of Raman thermograph or thermal reflectance techniques can potentially be improved with either a near-field or far-field nanoscopy technique.

Cross-References

- [Atomic Force Microscopy](#)
- [Carbon-Nanotubes](#)
- [Graphene](#)
- [Scanning Tunneling Microscopy](#)
- [Thermal Conductivity and Phonon Transport](#)

References

1. Williams, C.C., Wickramasinghe, H.K.: Scanning thermal profiler. *Appl. Phys. Lett.* **49**, 1587–1589 (1986)
2. Majumdar, A., Carrejo, J.P., Lai, J.: Thermal imaging using the atomic force microscope. *Appl. Phys. Lett.* **62**, 2501–2503 (1993)
3. Pollock, H.M., Hammiche, A.: Micro-thermal analysis: techniques and applications. *J. Phys. D Appl. Phys.* **34**, R23–R53 (2001)
4. Majumdar, A.: Scanning thermal microscopy. *Annu. Rev. Mater. Sci.* **29**, 505–585 (1999)
5. Aigouy, L., Tessier, G., Mortier, M., Charlot, B.: Scanning thermal imaging of microelectronic circuits with a fluorescent nanoprobe. *Appl. Phys. Lett.* **87**, 3 (2005)
6. Shi, L., Majumdar, A.: Thermal transport mechanisms at nanoscale point contacts. *J. Heat Trans-T. ASME.* **124**, 329–337 (2002)
7. Incropera, F.P., Dewitt, D.P., Bergman, T.L., Lavine, A.S.: *Fundamentals of Heat and Mass Transfer*. Wiley, Hoboken (2007)
8. Chen, G.: *Nanoscale Energy Transport and Conversion: A Parallel Treatment of Electrons, Molecules, Phonons, and Photons*. Oxford University Press, New York (2005)
9. Chung, J., Kim, K., Hwang, G., Kwon, O., Jung, S., Lee, J., Lee, J.W., Kim, G.T.: Quantitative temperature measurement of an electrically heated carbon nanotube using the null-point method. *Rev. Sci. Instrum.* **81**, 5 (2010)
10. Jo, I., Hsu, I.-K., Lee, Y.J., Sadeghi, M.M., Kim, S., Cronin, S., Tutuc, E., Banerjee, S.K., Yao, Z., Shi, L.: Low-frequency acoustic phonon temperature distribution in electrically biased graphene. *Nano Lett.* (2010). doi:10.1021/nl102858c
11. Cahill, D.G., Goodson, K., Majumdar, A.: Thermometry and thermal transport in micro/nanoscale solid-state devices and structures. *J. Heat Trans-T. ASME.* **124**, 223–241 (2002)
12. Tien, C.L., Majumdar, A., Gerner, F.M.: *Microscale Energy Transport*. Taylor & Francis, Washington, DC (1998)
13. Berciaud, S., Han, M.Y., Mak, K.F., Brus, L.E., Kim, P., Heinz, T.F.: Electron and optical phonon temperatures in electrically biased graphene. *Phys. Rev. Lett.* **104**, 227401 (2010)
14. Chae, D.H., Krauss, B., von Klitzing, K., Smet, J.H.: Hot phonons in an electrically biased graphene constriction. *Nano Lett.* **10**, 466–471 (2010)
15. Nakabepu, O., Suzuki, T.: Microscale temperature measurement by scanning thermal microscopy. *J. Therm. Anal. Calorim.* **69**, 727–737 (2002)
16. Price, D.M., Reading, M., Hammiche, A., Pollock, H. M.: Micro-thermal analysis: scanning thermal microscopy and localised thermal analysis. *Int. J. Pharm.* **192**, 85–96 (1999)
17. Kwon, O., Shi, L., Majumdar, A.: Scanning thermal wave microscopy (STWM). *J. Heat Trans-T. ASME.* **125**, 156–163 (2003)
18. Dunn, R.C.: Near field scanning optical microscopy. *Chem. Rev.* **99**, 2891–2928 (1999)
19. Hell, S.W.: Far-field optical nanoscopy. *Science* **316**, 1153–1158 (2007)
20. De Wilde, Y., Formanek, F., Carminati, R., Gralak, B., Lemoine, P.A., Joulain, K., Mulet, J.P., Chen, Y., Greffet, J.J.: Thermal radiation scanning tunneling microscopy. *Nature* **444**, 740–743 (2006)

Scanning Thermal Profiler

- [Scanning Thermal Microscopy](#)

Scanning Tunneling Microscopy

Ada Della Pia and Giovanni Costantini
Department of Chemistry, The University of Warwick, Coventry, UK

Definition

A scanning tunneling microscope (STM) is a device for imaging surfaces with atomic resolution. In STM, a sharp metallic tip is scanned over a conductive sample at distances of a few Å while applying a voltage between them. The resulting tunneling current depends exponentially on the tip-sample separation and can be used for generating two-dimensional maps of the surface topography. The tunneling current also depends on the sample electronic density of states, thereby allowing to analyze the electronic properties of surfaces with sub-nm lateral resolution.

Overview and Definitions

If two electrodes are held a few Å apart and a bias voltage is applied between them, a current flows even though they are not in contact, due to the quantum mechanical process of electron tunneling. This current depends exponentially on the electrode separation, and even minute, subatomic variations produce measurable current changes. In 1981, Gerd Binnig and Heinrich Rohrer at IBM in Zürich realized that this phenomenon can be used to build a microscope with ultrahigh spatial resolution [1], if one of the electrodes is shaped as a sharp tip and is scanned across the surface of the other (Fig. 1). Moreover, since the tunneling current depends also on the electronic properties of the electrodes, this microscope has the ability to probe the electronic density of states of surfaces at the atomic scale. A few years later, Don Eigler at IBM in Almaden, showed that, due to the extremely localized interaction between tip and sample, it is also possible to use this instrument to manipulate individual atoms, to position them at arbitrary locations and therefore to build

artificial structures atom-by-atom [2]. This remarkable achievement brought to reality the visionary predictions made by Richard Feynman in his famous 1959 lecture “*There’s plenty of room at the bottom*” [3].

The construction of this instrument, dubbed the scanning tunneling microscope (STM), was awarded the 1986 Nobel Prize in Physics and has since then revolutionized contemporary science and technology. The STM has enabled individual atoms and molecules to be imaged, probed, and handled with an unprecedented precision, thereby essentially contributing to our current understanding of the world at the nanoscale. Together with its offspring, the atomic force microscope (AFM) [► AFM], the STM is considered as the main innovation behind the birth of nanotechnology.

This entry will start with a discussion of the physical principles and processes at the heart of STM in section “[Theory of Tunneling](#).” This will be followed by a description of the experimental setup and the technical requirements needed for actually operating such a microscope in section “[Experimental Setup](#).” Section “[STM Imaging](#)” is dedicated to the most frequent use of STM, namely imaging of surfaces, while section “[Scanning Tunneling Spectroscopy](#)” gives a brief account of the spectroscopic capabilities of this instrument. Finally, section “[Applications](#)” discusses several applications and possible uses of STM.

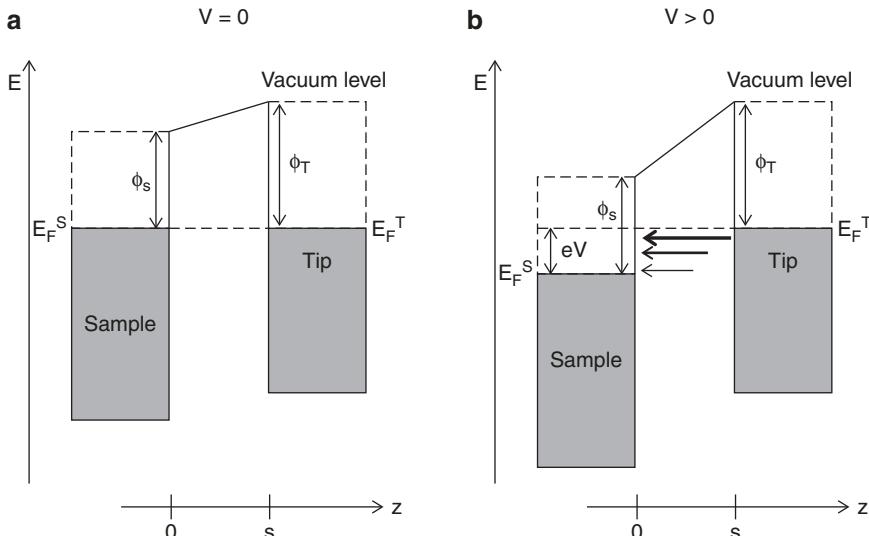
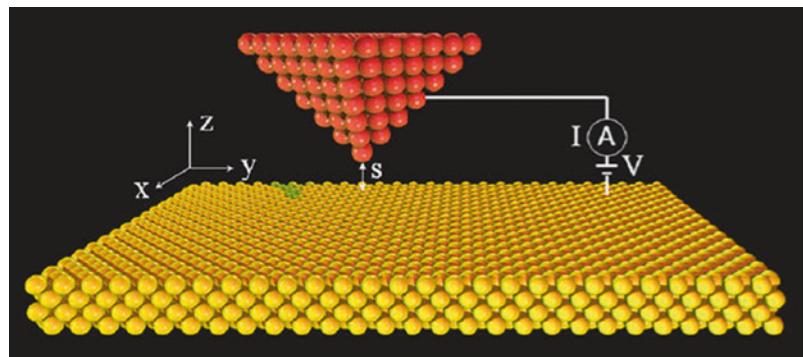
S

Theory of Tunneling

Figure 2 is a schematic representation of the energy landscape experienced by an electron when moving along the z axis of a metallic-substrate/insulator/metallic-tip tunneling junction. The following treatment can easily be extended to include also semiconducting tips or samples. Usually, the tip and the sample are not made of the same material and therefore have different work functions, ϕ_T and ϕ_S , respectively. At equilibrium, the two metals have a common Fermi level, resulting in an electric field being established across the gap region and in different local vacuum levels,

Scanning Tunneling Microscopy

Fig. 1 Schematic representation of a STM. Tip and sample are held at a distance s of a few Å and a bias voltage V is applied between them. The resulting tunneling current I is recorded while the tip is moved across the surface. The coordinate system is also shown



Scanning Tunneling Microscopy, Fig. 2 Energy potential perpendicular to the surface plane for an electron in a tip-vacuum-sample junction. z is the surface normal direction, and s is the tip-sample distance. The gray boxes represent the Fermi-Dirac distribution at 0 K. ϕ_{TS} and E_F^{TS} are the work functions and the Fermi levels of tip and sample, respectively. (a) Tip and sample in electrical

equilibrium: a trapezoidal potential barrier is created. (b) Positive sample voltage V : the electrons tunnel from occupied states of the tip into unoccupied states of the sample. The thickness and the length of the arrows indicate the exponentially decreasing probability that an electron with the corresponding energy tunnels through the barrier

depending on the difference $\phi_T - \phi_S$ (Fig. 2a). Since the work functions in metals are of the order of several eV, the potential in the gap region is typically much higher than the thermal energy kT and thus acts as a barrier for sample and tip electrons. A classical particle cannot penetrate into any region where the potential energy is greater than its total energy because this requires a negative kinetic energy. However, this is possible for electrons which, being quantum mechanical objects, are described by delocalized wave functions. This

phenomenon goes under the name of *quantum tunneling*. In an unpolarized tip-sample junction the electrons can tunnel from the tip to the sample and vice versa, but there is no net tunneling current. On the contrary, if a voltage V is applied between sample and tip, the Fermi level of the former is shifted by $-eV$ and a net tunneling current occurs, whose direction depends on the sign of V (Fig. 2b). Here the convention is adopted to take the tip as a reference since experimentally the voltage is often applied to the sample while the tip is grounded. If

V is the bias voltage, the energy for an electron in the sample will change by $-eV$, that is, it will decrease for positive values of V .

The tunneling current can be evaluated by following the time-dependent perturbation approach developed by Bardeen [4, 5]. The basic idea is to consider the isolated sample and tip as the unperturbed system described by the stationary Schrödinger equations:

$$(\mathcal{T} + \mathcal{U}_s)\psi_\mu = E_\mu\psi_\mu \quad (1)$$

and

$$(\mathcal{T} + \mathcal{U}_T)\chi_v = E_v\chi_v \quad (2)$$

where \mathcal{T} is the electron kinetic energy. The electron potentials \mathcal{U}_s and \mathcal{U}_T and the unperturbed wavefunctions ψ_μ and χ_v are nonzero only in the sample and in the tip, respectively. Based on this, it can be shown [5] that the transition probability per unit time $w_{\mu v}$ of an electron from the sample state ψ_μ to the tip state χ_v is given by Fermi's golden rule:

$$w_{\mu v} = \frac{2\pi}{\hbar} |M_{\mu v}|^2 \delta(E_v - E_\mu) \quad (3)$$

where the matrix element is:

$$M_{\mu v} = \int \chi_v^*(\vec{x}) \mathcal{U}_T(\vec{x}) \psi_\mu(\vec{x}) d^3x. \quad (4)$$

The δ function in Eq. 3 implies that the electrons can tunnel only between levels with equal energy, that is, (Eq. 3) accounts only for an *elastic tunneling process*. The case of an inelastic tunneling process will be considered in section “[Scanning Tunneling Spectroscopy](#).” The total current is obtained by summing $w_{\mu v}$ over all the possible tip and sample states and by multiplying this by the electron charge e . The sum over the states can be changed into an energy integral by considering the density of states (DOS) $\rho(E)$: $\sum \rightarrow 2 \int f(\varepsilon) \rho(\varepsilon) d\varepsilon$, where the factor 2 accounts for the spin degeneracy while f , the Fermi–Dirac distribution function, takes

into consideration Pauli's exclusion principle and the electronic state population at finite temperatures.

As a consequence, the total current can be written as:

$$I = \frac{4\pi e}{\hbar} \int_{-\infty}^{\infty} [f_T(E_F^T - eV + \varepsilon) - f_S(E_F^S + \varepsilon)] \times \rho_T(E_F^T - eV + \varepsilon) \rho_S(E_F^S + \varepsilon) |M|^2 d\varepsilon \quad (5)$$

where E_F is the Fermi energy and the indexes T and S refer to the tip and the sample, respectively. Equation 5 already accounts for the movement of electrons from the sample to the tip and vice versa.

Several approximations can be made to simplify Eq. 5 and to obtain a manageable analytical expression for I . If the thermal energy $k_B T \ll eV$, the Fermi–Dirac distributions can be approximated by step functions and the total current reduces to:

$$I = \frac{4\pi e}{\hbar} \int_0^{eV} \rho_T(E_F^T - eV + \varepsilon) \rho_S(E_F^S + \varepsilon) |M|^2 d\varepsilon \quad (6)$$

(Note that Eq. 6 is valid only for $V > 0$. For $V < 0$ the integrand remains identical but the integration limits become $-e|V|$ and 0). In this case, only electrons with an energy differing from E_F by less than eV can participate to the tunneling current. This can be directly seen in Fig. 2b for the case of positive sample bias: tip electrons whose energy is lower than $E_F^T - eV$ cannot move because of Pauli's exclusion principle, while there are no electrons at energies higher than E_F^T . The main problem in determining expression (Eq. 5) is, however, the calculation of the tunneling matrix elements M since this requires a knowledge of the sample and the tip wave functions, which can be very complicated. On the other hand, for relatively small bias voltages (in the ± 2 V range), Lang [6] showed that a satisfactory approximation of $|M|^2$ is given by a simple one-dimensional WKB tunneling probability. In the WKB approximation [7], the probability $D(\varepsilon)$

that an electron with energy ε tunnels through a potential barrier $U(z)$ of arbitrary shape is expressed as:

$$D(\varepsilon) = \exp\left\{-\frac{2}{\hbar}\int_0^s [2m(U(z) - \varepsilon)]^{\frac{1}{2}} dz\right\}. \quad (7)$$

This semiclassical approximation is applicable if ($\varepsilon \ll U$) which is generally satisfied in the case of metal samples where the work function is of the order of several eV. In order to obtain a simple analytical expression for D , the trapezoidal potential barrier of a biased tip-sample junction (see Fig. 2b) is further approximated with a square barrier of average height $\phi_{\text{eff}}(V) = (\phi_T + \phi_S + eV)/2$. By using this, the integral in Eq. 7 becomes:

$$D(\varepsilon, V, s) = \exp(-2ks) \quad (8)$$

where

$$k = \sqrt{\frac{2m}{\hbar^2}(\phi_{\text{eff}} - \varepsilon)}. \quad (9)$$

In order to evaluate k , it must be noted that electrons closest to the Fermi level experience the lowest potential barrier and are therefore characterized by an exponentially larger tunneling probability (see Fig. 2b). Thus, in a first approximation, it can be assumed that only these electrons contribute to the tunneling current which, for positive bias, is equivalent to set $\varepsilon \approx eV$ in Eq. 9. Moreover, if the bias is much smaller than the work functions, eV can be neglected, resulting in

$$k \cong \frac{\sqrt{m(\phi_T + \phi_S)}}{\hbar} = 5.1 \sqrt{\frac{\phi_T + \phi_S}{2}} \text{ nm}^{-1} \quad (10)$$

where the work functions are expressed in eV. Using typical numbers for metallic work functions, the numerical value of the inverse decay length $2 k$ in Eq. 8 becomes of the order of 20 nm^{-1} . Therefore, variations in s of 1 \AA correspond to one order of magnitude changes in the tunneling probability and, as a consequence, in the measured current. This very high sensitivity

provides the STM with a vertical resolution in the picometer regime. The lateral resolution of STM depends on how different points of the tip contribute to the total tunneling current. By considering a spherical tip shape with radius R , most of the current originates from the central position since this is closest to the surface. A point laterally displaced by Δx from the tip center is $\Delta z \approx \frac{\Delta x^2}{2R}$ further away from the substrate (higher order Δx terms are neglected in this evaluation). As a consequence, with respect to the tip center, the corresponding tunneling probability is reduced by a factor:

$$\exp\left(-2k \frac{\Delta x^2}{2R}\right). \quad (11)$$

By considering a tip radius $R \approx 1 \text{ nm}$, the current changes by one order of magnitude for variations $\Delta x = 3 \text{ \AA}$. The actual lateral resolution is typically smaller than this upper limit and can reach down to fractions of an \AA . Its specific value however depends on the precise shape of the tip which is unknown a priori. These values, together with the vertical resolution discussed above, lie at the basis of the STM atomic imaging capabilities.

Finally, if the tunneling probability (Eq. 8) is substituted for the tunneling matrix $|M|^2$ in Eq. 6, the total tunneling current can be expressed as:

$$I = \frac{4\pi e}{\hbar} \int_0^{eV} \rho_T(E_F^T - eV + \varepsilon) \rho_S(E_F^S + \varepsilon) e^{-2ks} d\varepsilon. \quad (12)$$

Therefore, for a fixed lateral position of the tip above the sample, the tunneling current I depends on the tip-sample distance s , the applied voltage V and the tip and sample density of states ρ_T and ρ_S , respectively.

Experimental Setup

As seen in the previous section, variations of 1 \AA in s induce changes in the tunneling probability of one order of magnitude. The exponential dependence in Eq. 8 is responsible for the ultimate spatial resolution of STM but places stringent

constraints on the precision by which s must be controlled, as well as on the suppression of vibrational noise and thermal drift. Moreover, typical tunneling currents are in the 0.01–10 nA range, requiring high gain and low noise electronic components. The following subsections are dedicated to a general overview of technologies and methods used to meet these specifications.

Scanner and Coarse Positioner

The extremely fine movements of the tip relative to the sample required for operating an STM are realized by using piezoelectric ([► Piezoresistivity](#)) ceramic actuators (*scanners*) which expand or retract depending on the voltage difference applied to their ends. In a first approximation, the voltage-expansion relation can be considered as linear with a proportionality factor (piezo constant) usually of few nanometer/Volt. The main requirements for a good scanner are: high mechanical resonance frequencies, so as to minimize noise vibrations in the frequency region where the feedback electronics operates (see section “[Electronics and Control System](#)”); high scan speeds; high spatial resolution; decoupling between x , y , and z motions; minimal hysteresis and creep; and low thermal drift. Although several types of STM scanner have been developed, including the bar or tube tripod, the unimorph disk and the bimorph [8], the most frequently used is a single piezoelectric tube whose outer surface is divided into four electrode sections of equal area. By applying opposite voltages between the inner electrode and opposite sections of the outer electrode, the tube bends and a lateral displacement is obtained. The z motion is realized by polarizing with the same voltage the inner electrode in respect to all four outer electrodes. By applying several hundred Volts to the scanner, lateral scan widths up to 10 μm and vertical ones up to 1 μm can be obtained, while retaining typical lateral and vertical resolutions of 0.1 and 0.01 nm, respectively.

While scanning is typically done by one individual piezoelectric element, larger displacements up to several millimeters are needed to bring the tip in close proximity to the sample, to move it to different regions of the surface or to exchange

samples or tips. These are achieved by mounting the scanner onto a coarse position device. Several designs have been developed to this aim including micrometric screws driven either manually or by a stepper motor, piezoelectric walkers like the louse used in the first STM [9] or the inch-worm [10], magnetic walkers where the movement is obtained by applying voltage pulses to a coil with a permanent magnet inside and piezoelectric driven stick-slip motors, as the Besocke-beetle [11] or the Pan motor [12].

Electronics and Control System

The voltages driving the piezoelectric actuators and their temporal succession and duration are generated by an electronic control system. The electronics are also used to bias the tunneling junction, to record the tunneling current and to generate the STM images. In most of the modern instruments, these tasks are digitally implemented by a computer interfaced with digital to analog (DAC) and analog to digital (ADC) converters. The tunneling current is amplified by a high gain I–V converter (10^8 – 10^{10} V/A) usually positioned in close proximity of the tip, so as to reduce possible sources of electronic interference. This signal is then acquired by an ADC and processed by the control system. DACs are used to apply the bias voltage (from a few mV to a few V) between tip and sample and, in conjunction with high voltage amplifiers, to polarize the piezo elements. A feedback loop is integrated into the control system and is activated during the frequently used *constant current* imaging mode (see section “[STM Imaging](#)”). By acting on the z motion of the scanner, the feedback varies s to keep the tunneling current constant. This is controlled by a proportional-integral and derivative (PID) filter whose parameters can be set by the operator. Finally, a lock-in amplifier is often used to improve the signal-to-noise ratio in scanning tunneling spectroscopy (STS) measurements (see section “[Scanning Tunneling Spectroscopy](#)”).

Tip

Sharp metal tips with a low aspect ratio are essential to optimize the resolution of the STM images and to minimize flexural vibrations of the tip,

respectively. Ideally, in order to obtain atomically resolved topographies and accurate spectroscopic measurements, the tip should be terminated by a single atom. In this case, because of the strong dependence on the tip-sample separation (see section “[Theory of Tunneling](#)”), most of the tunneling current would originate from this last atom, whose position and local DOS would precisely determine the tunneling conditions. In practice, however, it is almost impossible to determine the exact atomic configuration of the tip and the actual current is often due to a number of different atoms. This is still compatible with good tunneling conditions as long as these atoms are sufficiently localized (in order to avoid “multiple tip effects”) and their structural and chemical state remains constant during scanning.

The most commonly used methods to produce STM tips are to manually cut or to electrochemically etch thin wires of platinum-iridium and tungsten, respectively. These materials are chosen because of their hardness, in order to prevent tips becoming irreversible damaged after an accidental crash. Other metallic elements and even semiconductor materials have been used as tips for specific STM applications. Due to their chemical inertness, Pt-Ir tips are often used to scan in air on atomically flat surfaces without the need of any further processing. However, they typically have inconsistent radii, while etched W tips are characterized by a more reproducible shape. These latter have the drawback that a surface oxide up to 20 nm thick is formed during etching or exposure to air. For this reason, W tips are mostly used in ultrahigh vacuum (UHV) where the oxide layer can be removed through ion sputtering and annealing cycles. Prior to use, tips are often checked by optical microscopy, scanning electron microscopy ([► SEM](#)), and field ion microscopy or transmission electron microscopy ([► TEM](#)). The quality of a tip can be further improved during scanning by using “*tip forming*” procedures, including pulsing and controlled crashing into metal surfaces. These processes work because the desorption of adsorbed molecules or the coating with atoms of the metallic substrate can produce a more stable tip apex. If STM is performed in polar liquids ([► EC-STM](#)), electrochemical processes might generate Faradaic

or non-Faradaic currents which can be of the same order of magnitude or even larger than the tunneling current. In order to minimize these effects, the tip, except for its very apex, must be coated with an insulating material.

Vibration Isolation

A low level of mechanical noise is an essential requirement for any type of scanning probe microscopy. For this reason, the core of a STM, where the tip-sample junction is located, is always equipped with one or several types of vibration damping systems. These can be stacks of metal plates separated by elastic spacers, suspension springs, or eddy current dampers composed of copper elements and permanent magnets. The low-frequency components of mechanical noise (<10 Hz), which are the most difficult to eliminate, are minimized by building a small and rigid STM with a high resonance frequency. Depending on the overall size and weight of the microscope, further noise damping strategies can be adopted. Smaller, typically ambient conditions STMs can be placed on metal or granite slabs suspended by springs or bungee cords or floating on pneumatic isolators. Sometimes, piezo-driven, feedback-controlled active vibration suppressors are also combined with passive systems. Larger versions of pneumatic isolators and active damping are often used to float the frames and the chambers of big UHV STMs. The laboratory where a STM instrument is located also plays an essential role for its performance. Ground floor rooms are always preferred since they minimize low-frequency natural building oscillations, which can be very difficult to counteract. High-resolution instruments are sometimes placed on large concrete blocks which are separated from the rest of the laboratory floor and rest either on a sand bed, an elastomer barrier or on second-stage pneumatic isolators. Moreover, they are also often surrounded by an acoustically insulating box. All these systems essentially act as low-pass mechanical filters whose effectiveness improves with decreasing cutoff frequencies, that is, with increasing mass and decreasing rigidity. For this reason, the body of a STM is typically a relatively heavy block of metal and the frames, slabs, and

vacuum chambers supporting or containing the microscope often have a considerable weight.

Setups for Different Environments and Temperatures

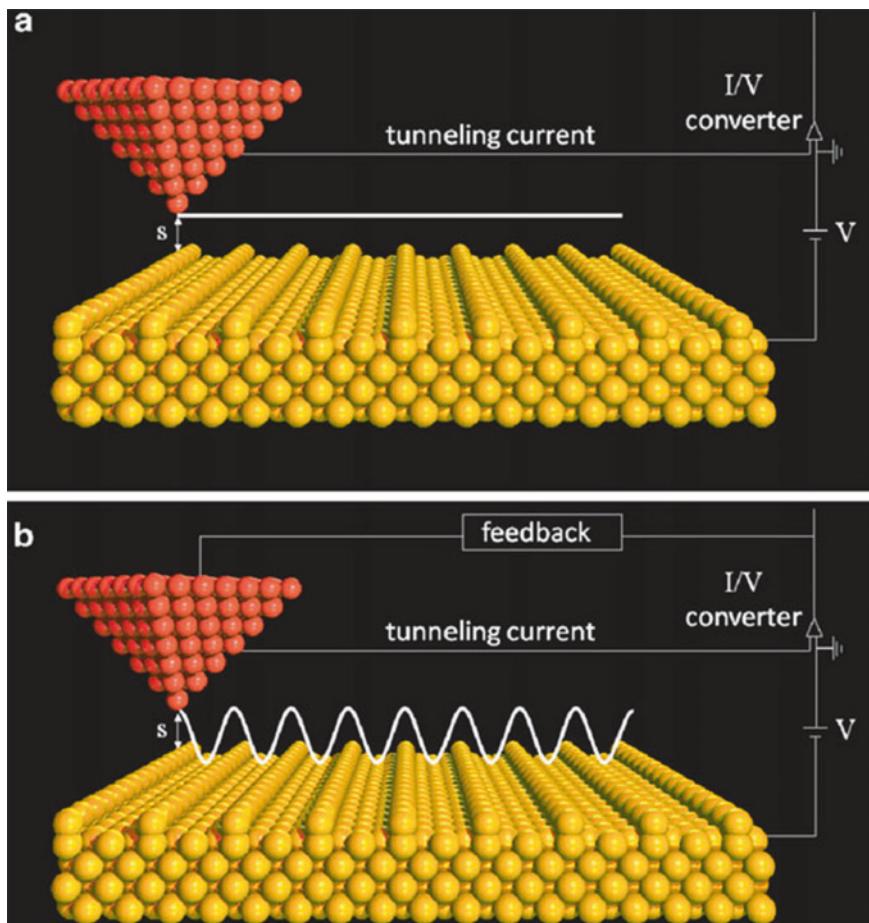
Different types of STMs have been developed that can operate in various environments such as air, inert atmosphere (N_2 , Ar), vacuum, high pressure, liquid, or in an electrochemical cell. The core of the different instruments is essentially the same, although the experimental chambers and setups in which they are located can vary substantially. Ambient condition STMs are typically quite compact and rigid and do not need elaborated anti-vibrational mechanisms. On the other hand, since sound waves represent a major problem, atmospheric pressure STMs are usually contained in an acoustic enclosure. A STM operating in vacuum must be hosted in a chamber with vibration-free pumps (typically ionic pumps for UHV) and must be equipped with sophisticated sample and tip manipulation mechanisms. Such systems often also have an *in situ* surface preparation stage allowing the handling of samples without air exposure.

STM can be performed at high pressures (1–30 bar) by installing the microscope head into gas manifolds under conditions similar to those used in industrial catalytic processes. Also in this case, sample and tip manipulation and preparation stages are mandatory parts of the system. Since these types of studies are typically performed at elevated temperatures (up to 600 K) and in the presence of highly reactive gases, the metallic parts of the STM scanner and of the chamber are often gold plated, the volume of the STM chamber is kept as small as possible and the tip material is chosen to be inert toward the gases [13]. Moreover, low voltages are used for polarizing the piezos in order to avoid gas discharges at intermediate pressures (10^{-3} –10 mbar) and shields are added to protect the STM from the deposition of conductive materials which could create electrical shorts.

STM at the liquid/solid interface and electrochemical STM (EC-STM) (► EC-STM) need the tip and sample to be inside a liquid cell which, in turn, may be placed in a humidity-controlled atmosphere. In the case of low vapor pressure liquids, the STM can be simply operated under

ambient conditions by dipping the tip into a liquid droplet deposited on the sample. A special coating must be applied to the tip when working with polar liquids (see section “Tip”).

STM can also be performed at different temperatures (in vacuum or controlled atmosphere chambers): variable temperature STM (VT-STM) able to cover the 5–700 K range, low temperature STM (LT-STM) operating at 77 K or 5 K and even milli-Kelvin STM instruments are currently available. A VT-STM is typically used to study thermally activated processes such as diffusion and growth, phase transitions, etc. These systems have sample heating and cooling stages which can be operated in a combined way so as to achieve a very precise temperature stabilization. Resistive heating is normally employed to increase the temperature, while both flow and bath cryostats with liquid nitrogen or helium as cryogenic fluids are used to reduce it. Continuous flow cryostats offer a high flexibility in temperature but are characterized by lower thermal stability, by inherent mechanical vibrations and do not easily attain temperatures below 20 K. Bath cryostats are more stable, are able to reach lower temperatures but are often also much bulkier (e.g., in order to limit the He consumption rate, a liquid He cryostat is actually a double-stage cryostat with an outer liquid nitrogen mantle). For most of these instruments the variable temperature capabilities refer to the possibility of choosing different (fixed) temperatures at which the microscope is run. However, few systems endowed with specific position tracking and drift compensating capabilities allow a “true” variable temperature operation where the same surface area can be imaged with atomic resolution while its temperature is changed. LT-STMs are operated at a fixed temperature and are typically inserted inside double stage cryostats which significantly complicates the tip and sample access. However, these instruments are extremely stable with a very low thermal drift and are therefore the best choice for STS and manipulation experiments (see sections “Scanning Tunneling Spectroscopy” and “Applications”). Milli-Kelvin STMs enable temperatures to be reached where extremely interesting magnetic, quantum Hall physics and superconductivity phenomena occur. Moreover, the thermal broadening of electronic



Scanning Tunneling Microscopy, Fig. 3 Schematic representation of (a) the constant height and (b) the constant current imaging modes, respectively. The thick lines represent the trajectory followed by the tip

features is strongly reduced, which is required for high-resolution measurements. These systems operate based on the evaporative cooling of liquid ^3He to temperatures of about 300 mK or liquid ^3He and ^4He mixtures below 10 mK. The STM heads can be further placed inside large-bore superconducting magnets (up to 15 T), enabling the low temperature and high magnetic field conditions necessary to access superconductive phase transitions or to detect single spin flip processes.

STM Imaging

STM images are generated by recording the tunneling current as a function of the tip position

while the tip is scanned across the sample surface. This can be done in two different ways which define the two main STM imaging modes:

- *Constant height mode.* The z section of the piezo scanner is kept fixed while the tip is moved over the substrate at a constant bias voltage (Fig. 3a). Variations of the tip-sample distance due to the surface topography produce a corresponding variation of the tunneling current which is recorded point-by-point and used to build the STM gray-level image. This mode is employed only in small areas of extremely flat surfaces, where the probability of crashing into protrusions such as steps or defects is relatively

- small. Very high scanning speeds can be used because of the absence of a feedback control.
- *Constant current mode.* While the x and y sections of the piezo scanner are used to laterally move the tip across the surface, the z section is driven by the electronic feedback so as to maintain a constant tunneling current (Fig. 3b). The corresponding z -voltage applied to the scanner (feedback signal) is recorded point-by-point and used to build the STM gray-level image. This mode can be employed for any type of surface topography and is therefore the most frequently used.

Since the constant height mode is applied to atomically flat surfaces with sub-Å height variations, the exponential $I-s$ relation derived from Eq. 12 can be approximated by a linear dependence. As a consequence, constant height STM images are a good representation of flat surfaces. On the other hand, for less planar substrates, one must use the constant current mode which directly reproduces the surface height due to the linear voltage-extension relation of piezoelectric materials. However, even constant current STM images are a reliable representation of the “true” surface topography only if the sample local DOS does not vary across the scanned area. If this is not the case, a constant current profile corresponds to a complex convolution of topographical and electronic features (see Eq. 12) which can be particularly relevant for surfaces covered with molecular adsorbates.

Scanning Tunneling Spectroscopy

Besides complicating the interpretation of STM images, the dependence of the tunneling current on the sample DOS also offers the unique opportunity of probing the electronic characteristics of surfaces with sub-nm spacial resolution. Having fixed the tip lateral position, the tunneling current I is a function of the applied bias voltage V and the tip-sample separation s only, the precise relation being established by Eq. 12. In a STS experiment, the relation between two of these three parameters is measured while the remaining one is kept

constant (STS). $I(V)$ spectroscopy, where the tunneling current is measured as a function of the bias voltage for a constant tip-sample separation, is the most widely used technique because it provides indications about the DOS of the sample.

Due to the spatial localization of the tunneling current (see section “[Theory of Tunneling](#)”), STS enables the characterization of the electronic properties of individual atoms and molecules in relation to their structure, bonding and local environment. Moreover, STS can also be used to create 2D maps of the sample DOS with sub-nm resolution. Such measurements are particularly interesting for quantum confined electronic systems (e.g., quantum dots or quantum corrals) or for determining the shape of molecular orbitals [14] (*wavefunction mapping*). By changing the polarity of the bias voltage, STS gives access to both the occupied and the unoccupied states of the sample. In this sense, it is often considered as complementary to ultraviolet photoemission spectroscopy (UPS), inverse photoemission spectroscopy (IPS) and electron energy loss spectroscopy (EELS), where the signal is averaged over a large area of the surface (between 0.1 and 2 mm in diameter). On the other hand, STS does not provide direct chemical information and tip artifacts can strongly influence the spectroscopic data.

So far we have assumed that electrons conserve their energy during the tunneling process (see Eq. 3). However, electrons can also tunnel inelastically between the tip and the sample by exchanging energy and inducing the excitation of vibrational modes, spin-flips, magnons, plasmons, excitons, etc. These extra tunneling channels become available only above specific voltage thresholds since only beyond these values a part of the electron energy can be converted into the excitation. The additional inelastic pathways increase the overall tunneling probability and therefore show up as discrete step-like features in the tunneling conductivity or as slope changes in $I(V)$ curves. This technique is called inelastic electron tunneling spectroscopy (IETS) and benefits from the same spatial resolution as STM and STS. IETS has been used to measure vibrational modes of individual molecules, spin excitations of single magnetic atoms, collective

plasmon excitations in 2D materials and magnons in ferromagnets.

A different way of detecting tunneling-induced molecular vibrations by means of a STM is to rely on their coupling with dynamical processes such as molecular motions. In particular, by measuring the frequency of molecular hopping events as a function of the applied bias voltage, it is possible to create so-called *action spectra* which reflect the vibrational spectrum of an individual molecule in a quantitative manner [15]. Optical excitations can also be revealed in an alternative way by coupling the STM with a photon detection system able to collect and analyze the luminescence stimulated by inelastically tunneling electrons [16]. Such a setup has been used to characterize plasmon emission from metallic surfaces and luminescence from semiconductor quantum structures and adsorbed molecules.

Applications

Since the first STM images of the surfaces of CaIrSn₄ and Au [1] were published back in 1982, STM has been used to analyze a wide range of materials: clean and adsorbate covered metal surfaces, semiconductors, superconductors, thin insulating layers, small and large organic molecules, individual atoms, liquid–solid interfaces, magnetic layers and surfaces, quasicrystals, polymers, biomolecules, nanoclusters, and carbon nanotubes. Imaging is the most frequent application of STM used to determine the structural properties of substrates and their reconstructions, the presence of defects, sites of adsorption for adatoms and molecules and the symmetry and periodicity of adsorbate superstructures. Nevertheless, right from the beginning, it became clear that the ultimate spatial resolution of STM, in combination with its dependence on the electronic properties of tip and sample, could allow a much wider range of applications of this instrument. These include the characterization of surface electronic, vibrational, optical, and magnetic properties, the measurement of single molecule conductivities, and the study of dynamic processes. In the following, we will only touch

upon some of the most frequent applications, without any presumption of being exhaustive.

Equation 12 shows that the tunneling current depends on the sample DOS close to the Fermi energy E_F . As a consequence, at a typical bias of a few Volts, it should be possible to image conductors, superconductors and small-gap or doped semiconductors but not molecules and insulating materials due to the vanishing DOS in the probed energy range (for most molecules the highest occupied and the lowest unoccupied orbitals are separated by an energy gap of several eV). However, the great majority of molecules adsorbed on metallic substrates can be easily imaged at moderate bias voltages. This is due to the formation of a metal-organic interface which can modify the molecular electronic properties leading to a broadening of the initial discrete energy levels, to a reduction of the gas-phase energy gap and even to the development of new states if covalent molecule-substrate bonds are established. All these effects contribute to the DOS at E_F and allow the imaging process. Regarding insulating materials, STM can only be done on films deposited onto conductive substrates if they are thin enough to allow the tunneling of electrons. These films are often used to electronically decouple organic adsorbates from metallic substrates.

The mechanism which allows the imaging of biomolecules such as DNA and proteins is currently still under debate [17]. As these molecules have a very large energy gap (5–7 eV) they can be considered as insulating materials and the current measured in STM experiments might be mediated by the thin water layer surrounding the molecules in air. Metalloproteins have also been imaged in their “natural” environment by using EC-STM (► EC-STM). Several reports have shown that when these redox active molecules are imaged under potentiostatic control, the tunneling current can be mediated by their metal redox-center, with enhanced conductivities measured for bias voltages close to the redox potential.

Although STM is a surface sensitive method, it can be also used to analyze buried interfaces and structures in cross-sectional STM (XSTM) [18]. The specific sample preparation in this technique requires brittle materials such as oxide samples or

semiconductor wafers. A cross section of the structure to be analyzed is prepared by cleaving the sample and positioning the STM tip onto the exposed edge. In this way, various physical properties can be probed, including the morphology and abruptness of buried nanostructures and interfaces, the alloying in epitaxial layers, the spatial distribution of dopants and their electronic configuration and the band offsets in semiconductor heterojunctions. XSTM has also been used to study, in real time, the changes occurring in semiconductor quantum well laser devices under operating conditions.

STM can further be employed for tracking dynamic surface processes, provided that the corresponding characteristic times are longer than the acquisition time. By choosing optimized designs for the piezo scanners and the electronic feedback, video-rate instruments have been developed able to record several tens of images per second and thereby to follow mobility and assembly processes in real time [19, 20].

When a tip with spin polarized electrons is used in STM, besides the parameters already indicated in Eq. 12, the local sample magnetization also influences the tunneling current. In fact, due to the different density of states at E_F of “spin-up” and “spin-down” electrons in magnetic materials, a spin polarized tip causes a tunnel magnetoresistance effect which results in a further contrast mechanism. This technique, called spin-polarized STM (SP-STM), has been used both in the presence and absence of an external magnetic field for detecting magnetic domain structures and boundaries in ferro- and antiferromagnetic materials, visualizing atomic-scale spin structures and determining spatially resolved spin-dependent DOS. An essential aspect of SP-STM is the ability to control the magnetization direction of the tip which can be achieved by evaporating different types of ferromagnetic or antiferromagnetic thin films on nonmagnetic tips. This technique is preferred to the use of bulk magnetic tips since it reduces the magnetic stray fields which can significantly modify the sample magnetization [21].

The ability of STM to identify and address individual nano-objects has been used to measure

the conductivity of single molecules absorbed on metal surfaces. While the tip is approached to the molecule of interest at constant bias voltage, the current flowing in the junction can be measured, thereby generating an $I(s)$ curve. Alternatively, $I(V)$ curves can be recorded at different s values. Since tip and substrate act as electrodes, both methods enable information to be obtained about the conductance of the individual molecule embedded in the junction. These measurements are often complemented by IETS experiments in the same configuration. IETS might in fact help to determine the arrangement and the coupling of the junction, which has a significant influence on the electronic and structural properties of the molecule. Single molecule STM conductance experiments represent an important source of information for understanding mechanisms of electron transport in organic molecules with applications in organic electronics and photovoltaics. They complement narrow gap electrode and break junction techniques, having the significant advantage of a highly localized electrode which allows to address and characterize individual molecules.

A similar type of application, although typically not aimed at individual molecules, is at the basis of the four point probe STM, where four STM tips, in addition to imaging, are used for local four point electric conduction measurements. A scanning electron microscope is installed above the STM enabling the positioning of the tips on the contact. The purpose of such very complex instruments is to measure the charge transport through individual nanoelectronic components (in particular self-assembled ones) and to correlate this information with a local high-resolution structural characterization.

S

Tip-Induced Modification

Besides being an extraordinary instrument for the characterization of structural, electronic, vibrational, optical, and magnetic properties of surfaces with subnanometer resolution, STM has also developed as a tool to modify and nanoengineer matter at the single molecule and atom scale.

By decreasing the distance between the tip and the sample in a controlled way, indentations can

be produced in the substrate with lateral sizes down to a few nm. Nanolithography can also be performed by tunneling electrons into a layer of e-beam photoresist (► [SU-8 Photoresist](#)), thereby reaching a better resolution compared to standard electron beam lithography (EBL). Many other STM-based nanopatterning and nanofabrication techniques have been developed based on a number of physical and chemical principles including anodic oxidation, field evaporation, selective chemical vapor deposition, selective molecular desorption, electron-beam induced effects, and mechanical contact. All these methods exploit the extreme lateral localization of the tunneling current and can be applied in air, liquids and vacuum.

However, the nanotechnological application that gained most attention is the ability to manipulate individual atoms and molecules on a substrate. This is possible due to a controlled use of tip-particle forces and is typically done in UHV and at low temperatures. The first atomic manipulation experiment was performed by Eigler and Schweizer in 1989 [2]. This phenomenal result fulfilled Richard Feynman's prophecy that "ultimately-in the great future-we can arrange the atoms the way we want; the very atoms, all the way down!" [3].

During a lateral manipulation experiment, the tip is first placed above the particle to be moved (for example an atom) and the tunneling current is increased while keeping a constant voltage. This results in a movement of the tip toward the atom, see Eq. 12. If their separation is reduced below 0.5 nm, Van der Waals forces start to come into play together with attractive and repulsive chemical interactions. When these forces equal the diffusion energy barrier, a lateral displacement of the tip can induce a movement of the atom parallel to the surface. After the desired final position is reached, the tip is retracted by reducing the tunneling current to the initial value, leaving the atom in the selected place. Depending on the tip-particle distance and therefore on the strength and nature of the interaction, different manipulation modes including pulling, pushing, and sliding [22] were identified and used to move different types of atoms and molecules.

Thanks to this technique, it was possible to fabricate artificial nanostructures such as the *quantum corral* [23] and to probe quantum mechanical effects like the quantum confinement of surface state electrons or the *quantum mirage*. Lateral STM manipulation has also been used to switch between different adsorption configurations and conformations of molecules on surfaces and to modify their electronic properties in a controlled way [24].

A further application of STM manipulation is the synthesis of new molecular species based on the ability of STM to form and break chemical bonds with atomic precision. Reactants are brought close together on the surface and the actual reaction is realized by applying a voltage pulse or by exciting vibrational modes through inelastically tunneling electrons. Examples of this technique include the dissociation of diatomic molecules, the Ullmann reaction, the isomerization of dichlorobenzene and the creation of metal-ligand complexes.

The STM tip has also been used to perform vertical manipulations of nanoparticles where an atom (or molecule) is deliberately transferred from the surface to the tip and vice versa by using the electric field generated by the bias voltage. In contrast to the lateral manipulation, here the bonds between the surface and the atom are broken and re-created [25]. By approaching the tip at distances of a few Å from the chosen particle chemical interactions are established that reduce the atom-surface binding energy. If a voltage pulse is applied under these conditions, the resulting electric field (of the order of 10^8 V/cm) can be enough to induce the particle desorption. The vertical manipulation technique has also been used as a means to increase the lateral resolution of STM. In fact, the controlled adsorption of a specific molecule onto the tip often makes it "sharper" and can add a chemical resolution capability if the DOS of the extra molecule acts as an "energy filter."

A related effect is exploited in the recently proposed scanning tunneling hydrogen microscopy (STHM) technique. In STHM, the experimental chamber is flooded with molecular hydrogen while the tip is scanned in constant

height mode at very close distances over the surface. H₂ can get trapped in the tip-sample junction and its rearrangement during scanning of the surface generates a new contrast mechanism based on the short-range Pauli repulsion. This is extremely sensitive to the total electron density, thereby endowing the STM with similar imaging capabilities to non-contact AFM (**► AFM, Noncontact Mode**) and making it able to resolve the inner structure of complex organic molecules [26].

Acknowledgments This work was supported by EPSRC (EP/D000165/1); A. Della Pia was funded through a WPRS scholarship of the University of Warwick. J. V. Macpherson, T. White, and B. Moreton are gratefully thanked for their critical reading of the manuscript.

Cross-References

- [AFM, Noncontact Mode](#)
- [Atomic Force Microscopy](#)
- [Electrochemical Scanning Tunneling Microscopy](#)
- [Electron Beam Lithography \(EBL\)](#)
- [Piezoresistivity](#)
- [Scanning Electron Microscopy](#)
- [SU-8 Photoresist](#)
- [Transmission Electron Microscopy](#)

References

1. Binnig, G., Rohrer, H., Gerber, C., Weibel, E.: Surface studies by scanning tunneling microscopy. *Phys. Rev. Lett.* **49**, 57 (1982)
2. Eigler, D.M., Schweizer, E.K.: Positioning single atoms with a scanning tunneling microscope. *Nature* **344**, 524 (1990)
3. Feynman, R.P.: There's plenty of room at the bottom: an invitation to enter a new field of physics. *Eng. Sci.* **23**, 22 (1960)
4. Bardeen, J.: Tunneling from a many-particle point of view. *Phys. Rev. Lett.* **6**, 57 (1961)
5. Gottlieb, A.D., Wesoloski, L.: Bardeen's tunnelling theory as applied to scanning tunnelling microscopy: a technical guide to the traditional interpretation. *Nanotechnology* **17**, R57 (2006)
6. Lang, N.D.: Spectroscopy of single atoms in the scanning tunneling microscope. *Phys. Rev. B* **34**, 5947 (1986)
7. Landau, L.D., Lifshitz, E.M.: Quantum Mechanics: Non-relativistic Theory. Pergamon Press, Oxford (1977)
8. Chen, C.J.: Introduction to Scanning Tunneling Microscopy. Oxford University Press, Oxford (2008)
9. Binnig, G., Rohrer, H.: Scanning tunneling microscope. *Helv. Phys. Acta* **55**, 726 (1982)
10. Okumura, A., Miyamura, K., Gohshi, Y.: The STM system constructed for analytical application. *J. Microsc.* **152**, 631 (1988)
11. Besocke, K.: An easily operable scanning tunneling microscope. *Surf. Sci.* **181**, 145 (1987)
12. Pan, S.H., Hudson, E.W., Davis, J.C.: ³He refrigerator based very low temperature scanning tunneling microscope. *Rev. Sci. Instrum.* **70**, 1459 (1999)
13. Laegsgaard, E., et al.: A high-pressure scanning tunneling microscope. *Rev. Sci. Instrum.* **72**, 3537 (2001)
14. Repp, J., Meyer, G., Stojkovic, S.M., Gourdon, A., Joachim, C.: Molecules on insulating films: scanning-tunneling microscopy imaging of individual molecular orbitals. *Phys. Rev. Lett.* **94**, 026803 (2005)
15. Sainoo, Y., et al.: Excitation of molecular vibrational modes with inelastic scanning tunneling microscopy processes: examination through action spectra of cis-2-butene on Pd(110). *Phys. Rev. Lett.* **95**, 246102 (2005)
16. Gimzewski, J.K., Reihl, B., Coombs, J.H., Schlittler, R.R.: Photon emission with the scanning tunneling microscope. *Z. Phys. B: Condens. Matter* **72**, 497 (1988)
17. Davis, J.J.: Molecular bioelectronics. *Philos. Trans. R. Soc. A* **361**, 2807 (2003)
18. Feenstra, R.M.: Cross-sectional scanning-tunneling-microscopy of III-V semiconductor structures. *Semicond. Sci. Technol.* **9**, 2157 (1994)
19. Rost, M.J., et al.: Scanning probe microscopes go video rate and beyond. *Rev. Sci. Instrum.* **76**, 053710 (2005)
20. Petersen, L., et al.: A fast-scanning, low- and variable-temperature scanning tunneling microscope. *Rev. Sci. Instrum.* **72**, 1438 (2001)
21. Wiesendanger, R.: Spin mapping at the nanoscale and atomic scale. *Rev. Mod. Phys.* **81**, 1495 (2009)
22. Bartels, L., et al.: Dynamics of electron-induced manipulation of individual CO molecules on Cu (111). *Phys. Rev. Lett.* **80**, 2004 (1998)
23. Crommie, M.F., Lutz, C.P., Eigler, D.M.: Confinement of electrons to quantum corrals on a metal-surface. *Science* **262**, 218 (1993)
24. Moresco, F., et al.: Conformational changes of single molecules induced by scanning tunneling microscopy manipulation: a route to molecular switching. *Phys. Rev. Lett.* **86**, 672 (2001)
25. Avouris, P.: Manipulation of matter at the atomic and molecular-levels. *Acc. Chem. Res.* **28**, 95 (1995)
26. Weiss, C., et al.: Imaging Pauli repulsion in scanning tunneling microscopy. *Phys. Rev. Lett.* **105**, 086103 (2010)

Scanning Tunneling Spectroscopy

Amadeo L. Vázquez de Parga and
Rodolfo Miranda

Department of Física de la Materia Condensada,
Universidad Autónoma de Madrid and Instituto
Madrileño de Estudios Avanzados en
Nanociencia (IMDEA-Nanociencia), Madrid,
Spain

Definition

Scanning tunneling spectroscopy (STS) is a technique that allows the study of the electronic structure of surfaces with atomic resolution.

Overview

Scanning tunneling microscopy (STM) was historically the second technique that could image individual atoms one by one. It was invented in 1981–1982 by Gerd Binnig and Heinrich Rohrer [1], long after the technique of field ion microscopy (FIM) developed in 1951 by Erwin Müller [2].

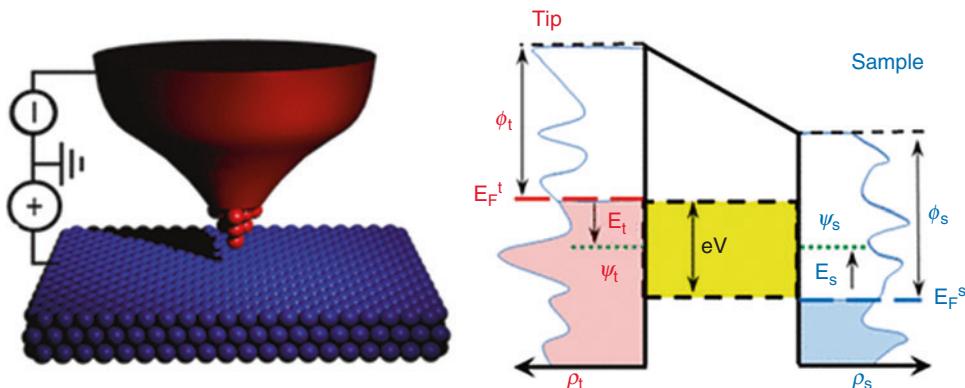
In STM a sharp tip probes the surface of interest by allowing electrons to tunnel quantum mechanically between the tip and the surface. Because such tunneling is extremely sensitive to the distance between tip and surface, one gets high resolution perpendicular to the surface. Assuming a constant density of states on the surface, when the STM tip is scanned over the sample surface while keeping the tunneling current constant, the tip movement depicts the surface topography, because the separation between the tip apex and the sample surface is always constant. It is worth noting that STM not only converts the spatial change in the tunneling current into a highly detailed topographic image of surfaces with constant density of states but also the tunneling current changes with the available surface electronic states. This dependence of the tunneling current on the surface electronic structure together with

the high spatial resolution of STM allows us to study the electronic structure of the surfaces with atomic resolution. The technique is known as scanning tunneling spectroscopy (STS).

Scanning Tunneling Spectroscopy Theory

Most theoretical treatments applied today to describe the tunneling process in an STM start from the formalism of the transfer Hamiltonian developed by Bardeen in 1961 [3] for the study of superconducting tunnel junctions. In this approach, the electronic structure and electron wave functions of both electrodes are calculated assuming no interaction between them and afterward the tunneling current is calculated [3]. Figure 1a shows the scheme of the tunnel junction in an STM where one of the electrodes is a tip. Figure 1b shows the energy diagram of the tunnel junction. In this diagram, the vertical axis represents energy. E_t and Ψ_t are the energy and wave function of the states of the electrode “tip” in the absence of electrode “sample.” E_s and Ψ_s are the energy and wave function of the states of the electrode “sample” in the absence of electrode “tip.” ϕ_t , ϕ_s , E_F^t , E_F^s , ρ_t , and ρ_s are the work functions, Fermi energies, and densities of states (DOS) of electrode “tip” and “sample,” respectively, and V is the voltage applied to electrode “sample.” When the distance between the electrodes is small enough, the overlap between their wave functions is significant, and the probability of electron transfer between the two electrodes by tunneling starts to be noticeable. In the absence of applied voltage, the Fermi levels of the two electrodes are aligned and no net tunneling current flows. However, by applying a voltage V , the Fermi levels move with respect to each other opening an energy window, eV , where electrons from one electrode can tunnel to the empty states of the other and, thus, the tunneling current starts to flow.

In 1983, Tersoff and Hamann applied the Bardeen’s formalism to the STM, replacing one of the electrodes by a point [4–6]. The tip was



Scanning Tunneling Spectroscopy, Fig. 1 *Left panel:* Configuration of a positively biased tunnel junction in an STM. The bias voltage is applied to the sample and the tunneling current is measured on the tip. *Right panel:* Energy diagram of an STM tunnel junction. In this diagram, the *vertical axis* represents energy and the *horizontal axes* distance between tip and sample and density of states.

shaped like an s orbital centered at the tip position and the calculated tunneling matrix elements proved to be proportional to the amplitude of the wave functions of the sample at the position of the tip. If the distance between tip and sample is not very large (few angstroms), the bias voltage small, and the temperature low, the tunneling current can be written as follows:

$$I \propto \int_0^{eV} \rho_s(\vec{r}_s, E) \rho_t(\vec{r}_s, E - eV) T(E, eV, d, \phi) dE \quad (1)$$

where \vec{r}_s is the tip position over the sample surface, d is the distance between tip and sample, and T is the transmission probability that depends on the energy of the states involved, the bias voltage applied between tip and sample, the distance between tip and sample, and the tunneling barrier height, which is related with the tip and surface work functions. This equation indicates that the tunneling process depends, for a given energy, on three interconnected parameters, i.e., the tunneling current I , the bias voltage V , and the tip sample separation d . Almost all attempts to explore these complex dependences of the tunneling current (and simultaneously extend the

E_t , Ψ_t , E_s , and Ψ_s are the energy and wave function of the states of the electrode tip and sample, respectively. ϕ_t , ϕ_s , E_F^t , E_F^s , ρ_t , and ρ_s are the work functions, Fermi energies, and densities of states (DOS) of electrode tip and sample, respectively, and V is the voltage applied to electrode sample

performance of STM) have been demonstrated by the late 1980s [7]. Scanning tunneling spectroscopy measures the relation between any two of while keeping fixed the third one. This gives three modes of spectroscopy measurements: (1) I-V curves, where the variation of the tunneling current with the bias voltage is measured for a fixed distance between tip and sample; (2) I-z curves, where the variation of the tunneling current with the distance between tip and sample is measured for a fixed bias voltage V ; and (3) V-z curves, where the variations in the tip sample distance are measured as function of the bias voltage for a fixed tunneling current.

In these three modes, energy conservation for the tunneling electrons is assumed. If electrons change their energy during the tunneling process by an inelastic process, the inelastic electron tunneling spectroscopy (IETS) mode is possible. Finally, if the STM tip and the sample are magnetic, the tunneling current depends on the relative orientation of the magnetization of both tip and sample. This mode is called spin-polarized scanning tunneling microscope (SP-STM) and allows the study of magnetic properties with atomic resolution. In the following, the five modes will be discussed in detail.

Scanning Tunneling Spectroscopy Modes

I-V Curves

I-V measurements are the most widely used spectroscopic technique in STM experiments. If the tunneling current (Eq. 1) is differentiated with respect to the bias voltage, the following expression is obtained:

$$\begin{aligned} \frac{dI}{dV} &\propto \rho_t(0)\rho_s(eV)T(eV, eV, d, \phi) \\ &+ \int_0^{eV} \rho_t(E - eV)\rho_s(E) \frac{dT(E, eV, d, \phi)}{dV} dE \\ &+ \int_0^{eV} \rho_s(E) \frac{d\rho_t(E - eV)}{dV} T(E, eV, d, \phi) dE \end{aligned}$$

Assuming a constant density of states for the tip, the third term is zero, but, it should be mentioned that, often, the tip electronic states have a strong influence in the STS spectra [8]. Another common simplification is to assume that the transmission coefficients are constant in the voltage range explored in the measurement. Then the second term also vanishes and the expression becomes

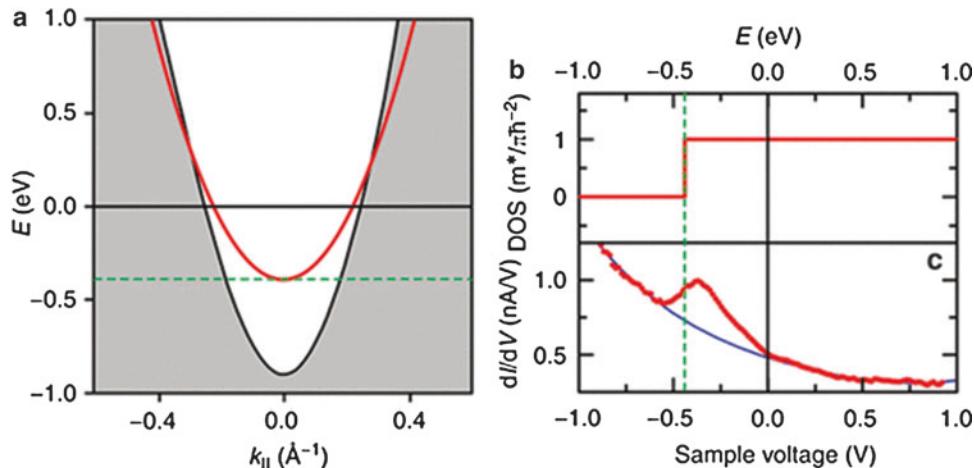
$$\frac{dI}{dV} \propto \rho_t(0)\rho_s(eV)T(eV, eV, d, \phi)$$

In general, the experimentally determined differential tunneling conductance is widely accepted as a good approximation to the DOS of the surface (modulated by the specific transmission of the barrier) at an energy value of eV , with V being the bias voltage applied between tip and sample.

Figure 2 shows an STS experiment performed on the Cu(111) surface in which the transmission probability (see Eq. 1) can be seen to depend on the energy parallel to the surface. Figure 2a shows the bulk band projection of Cu(111) along the $\overline{\Gamma M}$ direction of the surface Brillouin zone. Bulk bands are represented in gray and the projectional bandgap in white. The gray line corresponds to the dispersion relation of Shockley surface state of Cu (111). The surface state can be seen as a two-dimensional (2D) electron gas with the bottom of the band at -0.44 eV and an effective mass $m^* = 0.40 m_e$. Therefore, disregarding any contribution from the bulk electronic structure, the

LDOS expected around the Fermi level of Cu (111) is a step function centered at the bottom of the band. This step function is shown in Fig. 2b. Figure 2c shows an STS spectrum taken on the Cu (111) surface. The experimental data correspond to the dots. The spectrum has a peak at -0.38 eV superimposed on a background which decays with increasing energy. The bottom of the surface state band corresponds to the point halfway up the peak (dashed line in Fig. 2c). The line in the graphic is a fit to the background that reflects the contribution of the bulk states. As the bias voltage approaches the Fermi level from below, the k_{\parallel} of the accessible bulk states increases, so that they present a smaller effective perpendicular energy which means a smaller transmission probability and, thus, they contribute less to the spectrum. On the other hand, the peak at -0.38 eV corresponds to the sharp increase in the LDOS associated to the bottom of the surface state, superimposed on the background due to the bulk bands. Although the LDOS of a 2D electron gas is a step function to a constant value, the peak in the spectrum instead of staying constant for energies higher than the bottom of the band (-0.44 eV) decreases with increasing energy. This reduction in the signal with energy reflects the dispersion of the surface state, i.e., the fact that the k_{\parallel} of the surface state increases also when the energy increases. Accordingly, the transmission probability (and the signal in the spectrum) gets smaller.

Experimentally in order to record I-V curves, the distance between tip and sample has to be kept constant during the measurement time. This can be done in different ways. In one of them, the tip is placed at a desired position on the surface and at the desired distance from the surface. The distance between tip and sample is dictated by the values of the tunneling current and the bias voltage used in the topographic image. The feedback circuit, which keeps the tunneling current constant adjusting the tip sample distance, is disconnected and then the voltage V is ramped and the tunneling current is recorded over the desired bias voltage range. The dI/dV values are obtained by numerical differentiation of the I-V curves. If the measurement is repeated in every pixel of a topographic image, the method is called *current imaging*



Scanning Tunneling Spectroscopy, Fig. 2 (a) The bulk band projection of Cu(111) along the $\overline{T}\overline{M}$ direction of the surface Brillouin zone. Bulk bands are represented in gray, and in white the projectional bandgap. The gray line

corresponds to the dispersion relation of surface state of Cu(111). (b) Expected density of states for a 2D electron gas. (c) An STS spectrum taken on the surface Cu(111). The experimental data correspond to the dots

tunneling spectroscopy (CITS) and provides with a map of the spatial distribution of the LDOS on the surface.

Another method is to detect directly the dI/dV signal using a lock-in amplifier. In order to do that, a small high-frequency sinusoidal signal, $V_{\text{mod}} \sin(\omega t)$, is superimposed on the bias voltage between tip and sample. The modulation causes a sinusoidal response in the tunneling current, and the amplitude of the modulated current is sensitive to dI/dV . For a small applied sinusoidal signal, the modulated current can be Fourier decomposed on the applied modulation frequency ω :

$$I(V_{\text{bias}}, t) = I(V_{\text{bias}}) + \frac{dI(V_{\text{bias}})}{dV} V_{\text{mod}} \sin(\omega t) + \frac{d^2I(V_{\text{bias}})}{dV^2} \frac{V_{\text{mod}}^2}{4} \sin(2\omega t) + \dots \quad (2)$$

The first harmonic, which is proportional to the differential conductance (dI/dV), can be extracted by means of lock-in detection, and the spatial variation of the dI/dV signal can be mapped in certain area of the surface. During a constant current topographic image, the dI/dV signal is simultaneously recorded in each point of the image at a certain bias voltage. The result is a

map that reflects the LDOS of a surface area at a defined energy eV . Since the feedback loop is connected during the topographic image, the frequency of the modulated signal needs to be higher than the cutoff frequency of the feedback loop response in order to keep the tip distance constant during the acquisition of the data.

I-Z Curves

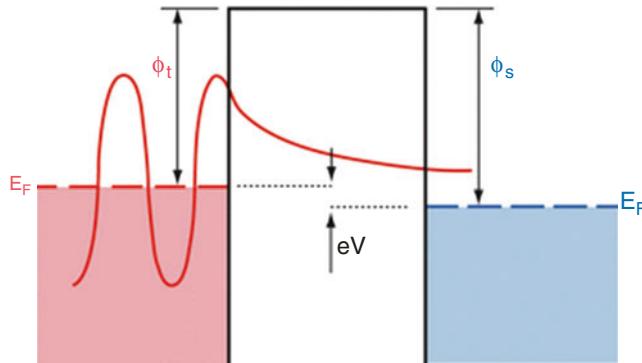
In the simplest theoretical treatment of the tunneling process for a metal-vacuum-metal tunnel junction [7] where the bias voltage is much smaller than the work function (assumed to be the same for both metals), the problem is reduced to a square potential barrier as shown in Fig. 3. The wave function that describes an electron in the tunneling barrier is given by:

$$\psi(z) = \psi(0)e^{-\kappa z}$$

where

$$\kappa = \frac{\sqrt{2m(\phi)}}{\hbar}$$

is the decay constant that describes the probability of finding the electron along the $+z$ direction and depends on the surface work function (ϕ).



Scanning Tunneling Spectroscopy, Fig. 3 One-dimensional metal-vacuum-metal tunnel junction. Both the sample, in red, and the tip, in blue, are modeled as semi-infinite pieces of free electron metals. The tunneling

The tunneling probability for a given distance, d , between tip and sample is

$$P \propto |\psi(0)|^2 e^{-2\kappa d}$$

and the tunneling current is proportional to the number of occupied states available in the energy interval defined by the bias voltage applied between tip and sample, and therefore, the tunneling current can be written as follows:

$$I \propto \sum_{E_n=E_F-eV}^{E_F} |\psi_n(0)|^2 e^{-2\kappa d}$$

If the bias voltage is small enough to consider the density of states constant, the equation can be written in terms of the local density of states (LDOS) at the Fermi level. At a position d and energy E , the LDOS of the sample can be expressed as

$$\rho_s(d, E) \equiv \frac{1}{\epsilon} \sum_{E_n=E-\epsilon}^E |\psi_n|^2$$

for small enough ϵ . The tunneling current in terms of the surface LDOS at the Fermi level is

$$I \propto V \rho_s(0, E_F) e^{-2\kappa d}$$

Assuming a typical value of 4 eV for the work function, $\kappa = 1.025 \text{ \AA}^{-1}$. From the expression

probability depends on the exponential decay of the electron wave function into the vacuum barrier. The bias voltage is small enough to consider a square barrier

above, the dependence of the logarithm of the tunneling current with respect to distance is a measure of the work function or, more precisely, of the tunneling barrier height. The corresponding expression is

$$\phi \approx 0.95 \left(\frac{d \ln I}{dz} \right)^2$$

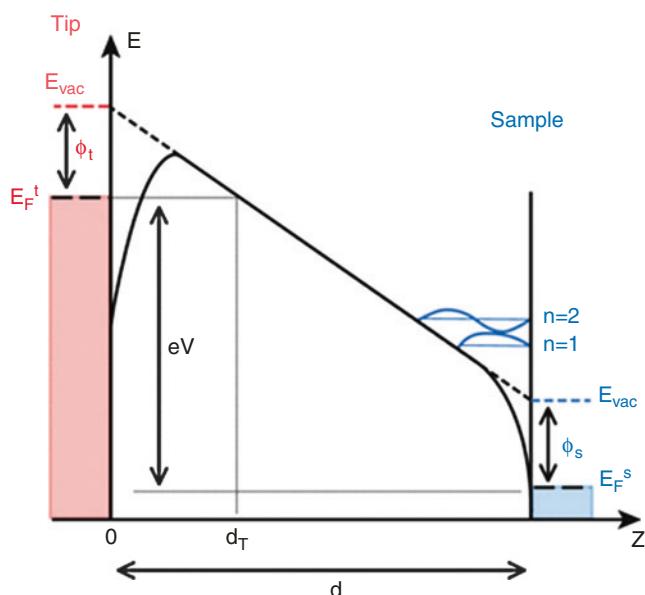
The measurement of the apparent barrier height can be carried out by approaching or retracting the tip from the sample and recording the tunneling current. In order to measure the spatial change in apparent barrier height, a small modulation in the separation between the tip and the sample is introduced at high frequency, and the modulated tunneling current is measured using a lock-in amplifier. This type of measurement gives the apparent barrier height at certain bias voltage and at certain distance from the surface. Measurements performed with different bias voltages or different tip sample distances may give different values for the apparent barrier height [7]. It is important to realize that the apparent barrier height is different from the work function in traditional surface science but is closely related; the apparent barrier height measures the spatial correlation of the overlap between the wave functions of the tip and the sample.

Z-V Curves

In Z-V measurements, the bias voltage is ramped at a fixed tunneling current, and the tip sample

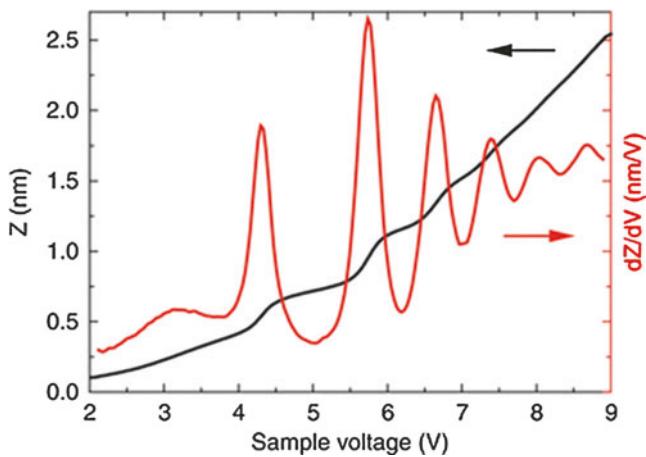
Scanning Tunneling Spectroscopy,

Fig. 4 Diagram of the tunnel junction when the positive bias voltage (eV) is larger than the work function of the sample. The electrons from the tip tunnel through a narrow (d_T) triangular potential barrier to be afterward trapped in a triangular potential well. The first two quantum well states are shown



Scanning Tunneling Spectroscopy,

Fig. 5 Black curve, tip displacement as function of the tunneling bias voltage. The field emission resonances appear as wiggles on the curve. Gray curve, dZ/dV , the field emission resonances appear as well-defined peaks that allow a precise determination of the energy



separation is constantly adjusted. When the applied voltage exceeds the sample or tip work function (depending on the sign of the bias voltage), there is a transition from the vacuum tunneling regime to the field emission regime. In the field emission regime, a triangular potential well is formed between tip and sample due to the bias voltage applied in the tunneling junction. In this triangular potential well, the existence of quantum well states leads to resonances in the electron transmission at certain energies, as illustrated in Fig. 4.

These transmission resonances show up as wiggles in the Z-V data (black curve in Fig. 5) and are closely related to the image states. Image states are unoccupied states bound by the classical image charge response of metallic surfaces and have a free electron-like dispersion parallel to the surface. The inverse dependence on distance from the surface of the image potential leads to a Rydberg-like series of states that converges to the continuum at the vacuum level (E_{vac}). Inverse photoemission studies of the image potential states have shown experimentally that the energy

position of the Rydberg series is tied to the local surface potential of the material. In STM, the electric field across the tunnel junction causes a Stark shift of these states, expanding the image state spectrum into a resonance spectrum associated with the triangular potential well (Fig. 5). Following the analysis performed by Gundlach in the 1960s [9], the resulting energy spectrum can be written as follows:

$$E_n = \phi + \alpha(n - 0.25)^{2/3} F^{2/3}$$

where ϕ is the surface work function, α is a constant, F is the electric field between tip and sample, and n is the quantum number of the states. These field emission resonances (FERs) were experimentally observed in field ion microscopy (FIM) by Jason [10] and with an STM by Binnig et al. [11] and since then have been used to chemically identify different transition metals on surfaces, to obtain atomic resolution on insulating surfaces (e.g., diamond), or to study local changes in the surface work function [12].

The experiments are typically performed with the feedback loop on to keep the current constant. The tip movement is recorded as a function of the bias voltage, and afterward, the curves are numerically differentiated to obtain the energy position of the field emission resonances, as shown in Fig. 5 (red curve).

Spin-Polarized Tunneling Spectroscopy

In 1975 Julliére [13] discovered spin-dependent tunneling between two planar ferromagnetic electrodes separated by an insulating tunnel barrier, which has become the basis for the development of magnetic random access memories and the spin-polarized version of the STM. In fact, the tunneling current between a magnetic sample and an STM tip (covered with a magnetic thin film) shows an asymmetry in the spin population. The magnitude of the tunneling conductance between two magnetic electrodes with directions of the respective magnetization differing by a certain angle depends on the cosine of this angle. Spin-polarized tunneling with an STM was observed in 1990s by Wiesendanger [14]. The spectroscopic mode of spin-polarized STM is

based on using the different intensity of certain features in differential conductance spectra as source of contrast to image magnetic domains and domain walls with atomic resolution [14].

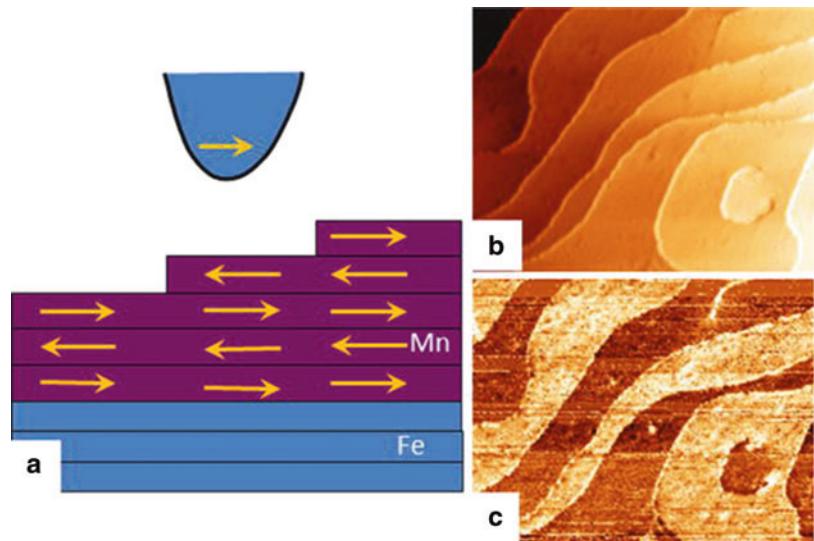
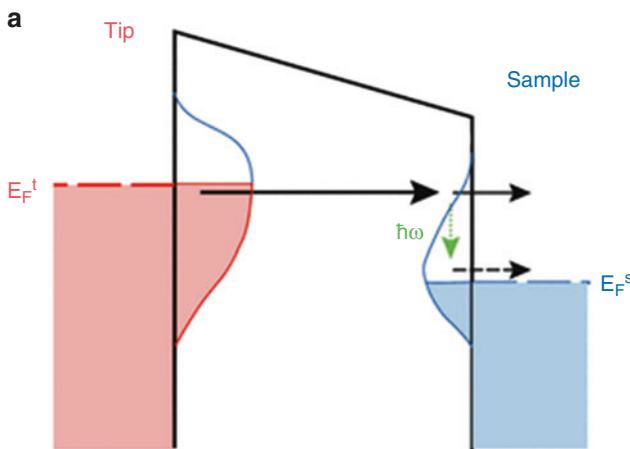
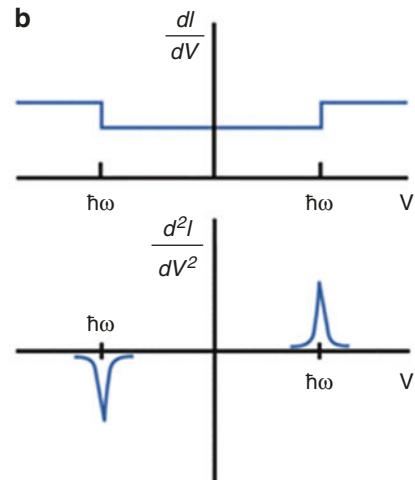
Magnetic domain observations in antiferromagnetic materials have been difficult in the past due to the limited number of experimental techniques that are sensitive to domain states in antiferromagnetic crystals. Another aspect that hampered the studies of antiferromagnetic materials was the limited spatial resolution of the available techniques. Mn and Cr crystals or thin films of these materials exhibit quite complex spin structures. The crystallographic structure of Mn thin films grown on Fe(001) is body-centered tetragonal. In this crystallographic structure, the Mn magnetic structure is layered antiferromagnetic. This means that the magnetization orientation rotates by 180° in every layer, as can be seen in Fig. 6a. The tip magnetization direction is constant in the experiment, and the magnetization direction of the sample rotates by 180° every time the tip crosses a step on the surface. The change in the magnitude of the tunneling conductance can be measured on every pixel of the topographic image, and an image of the magnetic domains is obtained, as can be seen in Fig. 6b, c. In those domains where the magnetization of tip and sample is aligned, the tunneling conductance is higher (brighter color), and on those areas where antiparallel, the tunneling conductance is lower (darker colors).

Vibrational Spectroscopy (Inelastic Electron Tunneling Spectroscopy)

In the previous sections, elastic tunnel processes, where the tunneling electrons do not change their energy, have been discussed. However, in certain cases, i.e., for molecules adsorbed on surfaces or samples with easy excitation of phonons, there is a small fraction of electrons that lose energy in the tunneling process [15]. For bias voltages larger than corresponding the quantum of vibration, $\hbar\omega$, a new tunneling channel, i.e., inelastic channel, opens up (as illustrated in Fig. 7a). The inelastic channel acts in addition to the elastic channel and increases slightly the differential conductance (dI/dV) of the junction (Fig. 7b upper panel).

Scanning Tunneling
Spectroscopy, Fig. 6

(a) Model showing the layered antiferromagnetic structure of Mn films grown on Fe (001). (b) STM topographic (100×78 nm) image of 6.5 mL of Mn grown on Fe (001). (c) Spatially resolved spectroscopic image measured simultaneously with the topography shown in (b) where the dI/dV signal shows low and high levels depending on the relative directions of magnetization of tip and surface terrace revealing the topological antiferromagnetic order of the Mn(001) surface

**a****b**

Scanning Tunneling Spectroscopy, Fig. 7 (a) Energy distance diagram of the tunneling processes with an applied bias voltage V . When eV is larger than the energy of the molecular vibration ($\hbar\omega$), empty final states at the Fermi level of the sample become accessible, and the

inelastic channel opens up. (b) The opening of the inelastic channel causes a sharp increase in the tunneling conductance (upper panel) or peaks in the second derivative (lower panel). The activation channel is symmetric with respect to the Fermi level

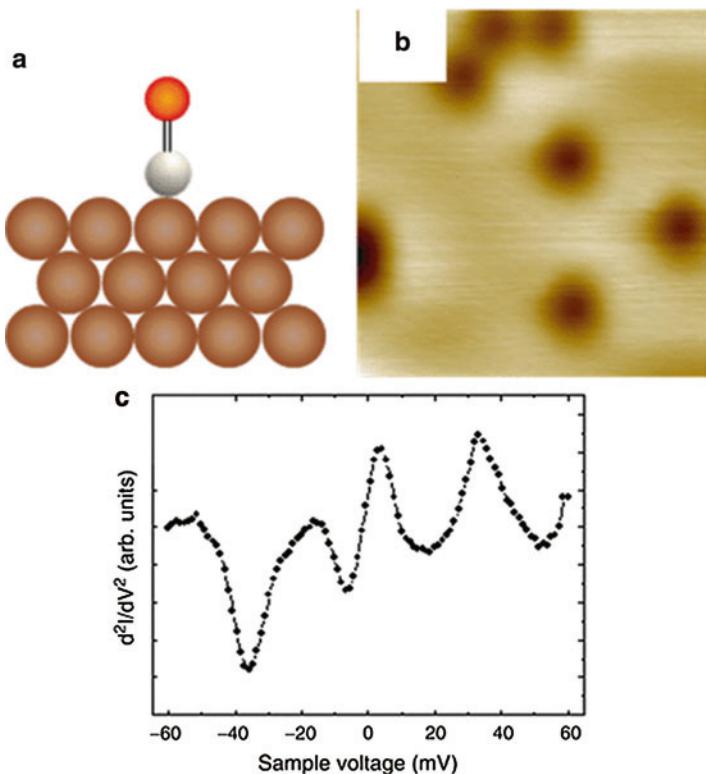
Although vibrations of molecules were detected in the 1960s by tunneling in extended tunneling junctions with insulating layer spray coated with molecules, the use of STM facilitates the acquisition of vibrational spectra in single molecules in well-characterized environments and was pioneered by Wilson Ho [15].

In practice the change in conductance is smaller than 10 % and can be detected only

under very severe conditions of stability of the tunnel junction and energy resolution (i.e., with the STM at low temperatures). The vibrational modes are detected as peaks in the second derivative of the tunneling current (Fig. 7b lower panel) measured by means of lock-in techniques. Equation 2 shows that the magnitude of the second harmonic of the tunneling current is proportional to d^2I/dV^2 . In practice, a small (≈ 3 mV) ac

Scanning Tunneling
Spectroscopy, Fig. 8

(a) Model of the adsorption of CO molecules on Cu(111). The molecule adsorbs perpendicular to the surface with the carbon (in gray) chemically bonded to the surface copper atoms. (b) STM topographic image measured using a tunneling current of 1 nA and a sample bias voltage of 0.25 V. (c) Vibrational spectra of CO on Cu(111): the peaks at ± 5 mV and ± 35 mV are due to excitation of the CO frustrated rotation and translation with respect to the Cu(111) surface



component is added to the bias voltage, the dc component of the bias voltage is scanned across the selected energy range, and the variations in d^2I/dV^2 are recorded. The width of the peaks is given by the Fermi energy distribution, i.e., the FWHM is $3.5k_B T$ (1.2 mV at 4 K).

A common observation in IETS spectra is that peaks at certain values of positive voltages appear as *dips* at opposite polarity. The symmetry position with respect to the zero bias of the features observed is a fingerprint of their inelastic origin. Differences in the density of states of the electrodes, however, may change the peak intensity. The observed symmetry implies that the inelastic processes are accessible for electron tunneling on both directions. Selection rules for which modes are detectable, unlike Raman or infrared spectroscopy, seem to depend on the symmetry of the molecular state involved in the tunneling process.

A well-studied system are CO molecules adsorbed on Cu(111). It is known that the CO molecules adsorb on the surface on top of the

copper atoms with the oxygen pointing toward the vacuum and the carbon chemically bonded to the copper surface, as shown in Fig. 8a. Figure 8b shows a topographic STM image of several CO molecules adsorbed on Cu(111) at 4.6 K measured with a tunneling current of 1 nA and a bias voltage of -0.25 V. With these values, the CO molecules are imaged as a round depression because the chemical bond between the molecule and the Cu(111) surface reduces the electron density around the Fermi level. In order to measure the vibrational spectra, the tip is positioned over the center of the CO molecule. With the feedback off, the sample bias voltage is ramped over the range of the vibrational peaks while a sinusoidal bias modulation is superimposed. The derivative of the conductance exhibits peaks at the molecular vibration energy. For CO on Cu(111), the vibration spectra are characterized by two features at about 5 and 35 meV, as can be seen in Fig. 8c. These peaks are assigned to the two degenerated transverse vibration modes: the frustrated translation and the frustrated rotation, respectively.

Cross-References

- Field Electron Emission from Nanomaterials
- Scanning Tunneling Microscopy
- Surface Electronic Structure

References

1. Binnig, G., Rohrer, H.: Scanning tunneling microscopy. *Helv. Phys. Acta* **55**, 726 (1982)
2. Müller, E.W.: Das Feldionenmikroskop. *Z. Phys* **31**, 136 (1951)
3. Bardeen, J.: Tunneling from many particle point of view. *Phys. Rev. Lett.* **6**, 57 (1961)
4. Tersoff, J., Hamann, D.R.: Theory and application for the scanning tunneling microscope. *Phys. Rev. Lett.* **50**, 1998 (1983)
5. Tersoff, J., Hamann, D.R.: Theory of the scanning tunneling microscope. *Phys. Rev. B* **31**, 805 (1985)
6. Lang, N.D.: Spectroscopy of single atoms in the scanning tunneling microscope. *Phys. Rev. B* **34**, 5947 (1986)
7. Chen, C.J.: Introduction to Scanning Tunneling Microscopy. Oxford University Press, New York (1993)
8. Vázquez de Parga, A.L., Hernán, O.S., Miranda, R., LeviYeyati, A., Mingo, N., Martín-Rodero, A., Flores, F.: Electron resonances in sharp tips and their role in tunneling spectroscopy. *Phys. Rev. Lett.* **80**, 357 (1998)
9. Gundlach, K.H.: Zu berechnung des tunnelstroms durch eine trapezförmige potentialstufe. *Solid State Electron.* **9**, 949 (1966)
10. Jason, A.J.: Field induced resonance states at a surface. *Phys. Rev.* **156**, 266 (1966)
11. Binnig, G., Frank, K.H., Fuchs, H., Garcia, N., Reihl, B., Rohrer, H., Salvan, F., Williams, A.R.: Tunneling spectroscopy and inverse photoemission: image and field states. *Phys. Rev. Lett.* **55**, 991 (1985)
12. Borca, B., Barja, S., Garnica, M., Sánchez-Portal, D., Silkin, V.M., Chulkov, E.V., Hermanns, C.F., Hinarejos, J.J., Vázquez de Parga, A.L., Arnau, A., Echenique, P.M., Miranda, R.: Potential energy landscape for hot electrons in periodically nanostructured graphene. *Phys. Rev. Lett.* **105**, 036804 (2010)
13. Julliere, M.: Tunneling between ferromagnetic films. *Phys. Lett.* **54A**, 225 (1971)
14. Wiesendanger, R.: Soin mapping at the nanoscale and atomic scale. *Rev. Mod. Phys.* **81**, 1495 (2009)
15. Person, B.N.J., Baratoff, A.: Inelastic electron tunneling from a metal tip: the contribution from resonant processes. *Phys. Rev. Lett.* **59**, 339 (1987)
16. Stipe, B.C., Rezaei, M.A., Ho, W.: Single molecule vibrational spectroscopy and microscopy. *Science* **280**, 1732 (1998)

Scanning X-Ray Diffraction Microscopy (SXDM)

- Selected Synchrotron Radiation Techniques

Scanning-Probe Lithography

- Dip-Pen Nanolithography

Scincus officinalis

- Friction-Reducing Sandfish Skin

Scincus scincus

- Friction-Reducing Sandfish Skin

Scrolled Nanostructure

- Nanorobotics for NEMS Using Helical Nanostructures

Selected Synchrotron Radiation Techniques

Antoine Barbier¹, Cristian Mocuta² and Rachid Belkhou²

¹CEA-Saclay, DSM/IRAMIS/SPCSI,
Gif-sur-Yvette, France

²Synchrotron SOLEIL, L'Orme des Merisiers,
Gif-sur-Yvette, France

Synonyms

Small angle X-ray scattering in grazing incidence geometry; Scanning X-ray diffraction microscopy (SXDM); Soft X-ray microscopy; Spectromicroscopy; X-ray diffraction with micron sized X-ray beams

Definition

Synchrotron radiation is produced when highly energetic charged particles are deviated in a magnetic field. The generated spectrum consists of a large energy range from infrared to gamma X-rays. Harder X-ray ranges require larger synchrotron storage rings. Recent rings offer simultaneously energy tunability for photons with high brightness, high photon fluxes, and low divergence [1]. Photon energy tunability and variable polarization provide chemical, electronic structure and/or magnetic sensitivity. The tunable penetration depth using glancing incident and/or exit scattering angles enables the study of buried interfaces. Importantly, X-ray techniques, especially at high photon energies, can be used with various sample environments including (and not limited to) high pressure, high temperatures, ultrahigh vacuum, liquids, and low temperatures. This unique combination of characteristics allowed developing genuine experimental methods and techniques particularly suitable to investigate micro- and nanostructures that are intended to complement laboratory experiments in order to obtain in-depth information on specific issues.

The use of synchrotron radiation has also some drawbacks. The available beam time is limited. Organic and biological samples are often unable to bear the highly intense X-ray beams. X-ray diffraction techniques measure scattered intensities, and the phase information is lost; unless the coherence of the synchrotron radiation or statistical methods are used the data interpretation requires fit procedures and the input of some models.

The goal of the present essay is here to exemplify some approaches proven to be relevant to investigate nanostructures of current interest. The interested reader may consider more general textbooks or reviews dedicated to synchrotron radiation techniques [2, 3]. Within the field of nanostructure investigations three techniques that are particularly illustrative were chosen: Grazing Incidence Small Angle X-ray Scattering (GISAXS), micro X-Ray Diffraction (μ -XRD), and X-ray Photo-Emission Electron Microscopy (X-PEEM). These techniques are essentially

nondestructive, noninvasive, and allow investigating surfaces and buried interfaces. GISAXS and μ -XRD allow studying samples in various environments (e.g., temperature, gas or liquid cells for “real” environments, external magnetic or electric fields, etc.). X-PEEM is more limited within this respect and requires good vacuum conditions in order to detect electrons but provides full field real space imaging capabilities.

Overview

X-ray scattering probes materials properties and, depending on the technique used, can be sensitive to chemical species, magnetic and/or crystallographic ordering, and even density fluctuations near surfaces and buried interfaces. The process can be elastic (i.e., photon energy conservative) or inelastic, specular or diffuse [4], and can be used in spectroscopies that provide chemical information about the sample composition. Photon energy tunability can be used to greatly enhance the chemical contrast between several chemical species (a specific contrast is obtained when the photon energy is tuned to a given electron level edge), i.e., anomalous x-ray scattering conditions [2, 3].

For the analysis of materials at the micro- and the nanoscale, well-established microscopy-based techniques can probe different properties at short – down to atomic – length scale. Techniques like the transmission electron microscopy (► TEM), scanning probe methods (scanning electron microscopy ► SEM, atomic force microscopy ► AFM, microphotoluminescence μ PL, etc.) have an undeniable contribution in understanding what is happening at the small scale for nanostructured materials. From some of their drawbacks, one could mention here their surface sensitivity, invasive sample preparation methods, presence of vacuum environment, or sensitivity to external fields (magnetic, electric). Alternative X-ray-based techniques can offer complementary insights and overcome some of these drawbacks. Some examples are detailed hereafter.

GISAXS is a powerful technique to study nanostructured surfaces, thin films, and assemblies of nano-objects, combining small-angle

X-ray scattering (SAXS) and the surface sensitivity of grazing incidence diffraction (GID) [5]. The method is fully nondestructive for materials science samples and gives access to morphological (e.g., size/shape, dispersion) statistical information of an assembly of nano-objects (separated by an average distance in the several 10–100 nm range), averaged over a large area (several square millimeters); it characterizes the “mean” object in reciprocal space. During a GISAXS experiment the incident photon energy is mainly chosen with respect to the reciprocal space resolution that is expected. For chemical contrast studies the photon energy may be tuned to the absorption edge of a given element. The pattern is conveniently observed on an area detector (2D camera) because at high photon energy the Ewald sphere region corresponding to these small scattering angles can reasonably be approximated by the plane tangent to the sphere.

Microbeam X-ray diffraction (**μ-XRD**) experiments use highly intense and very small (focused to micron or submicron sizes) X-ray beams in order to address locally the properties of matter by diffraction [4, 6]. The lateral resolution is essentially given by the size of the X-ray spot. XRD uses typically hard X-rays (~ 10 keV energy range). It is a scanning microscopy-like approach and consists of investigating a single structure at a time. The signal originating from individual micro/nanostructures is recorded by using local probe techniques, and a two-dimensional raster image of the sample is obtained. It can give access to crystalline structure, strain, defects, composition (by exploiting the variation of the lattice unit cell size with the composition), etc.; the strain can be measured in a simple experiment with an accuracy (10^{-4}) hard to achieve with other techniques. The resulting data (images or maps in the so-called reciprocal space, which is the Fourier transform space) are however usually not straightforward to visualize and understand (as compared to real space imaging techniques).

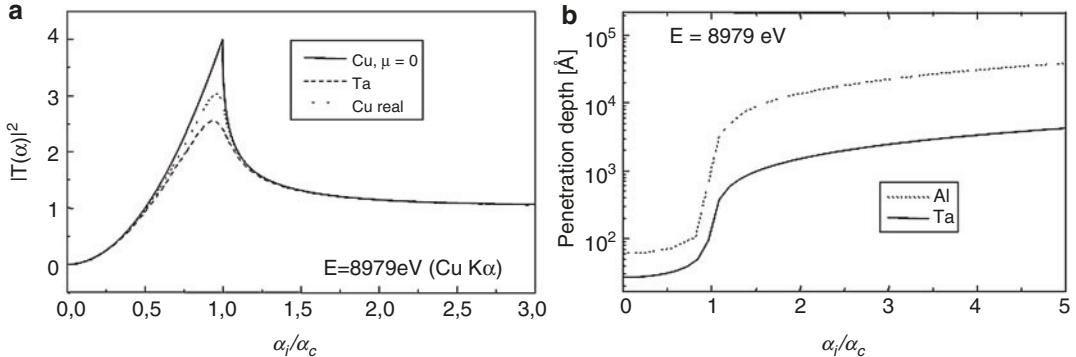
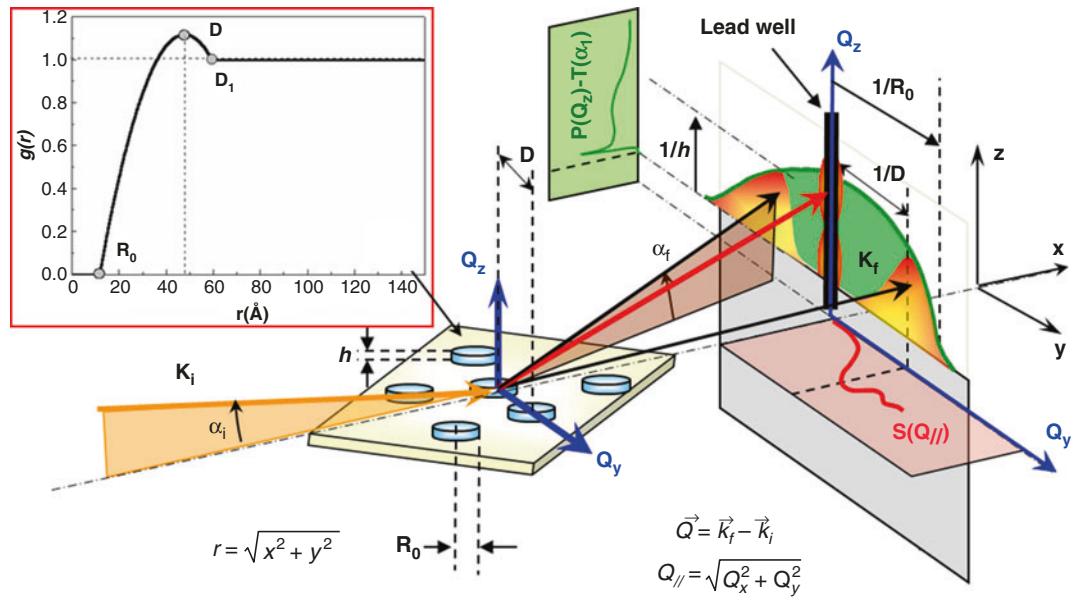
XPEEM (X-ray PhotoEmission Electron Microscopy) spectromicroscopy is a derivative of the classical PEEM/LEEM [7, 8] employing parallel imaging techniques making use of special electron optics with resolution in the 10 nm

ranges. If a photon energy just above the photothreshold is used, the photoelectron yield is mainly determined by the differences in the work function ϕ of the sample. The local variations of ϕ result in images with high contrast. This ultraviolet (UV)-PEEM mode of operation is ideally suited to study surface chemical reactions in real time. It allows accessing well-established techniques like Ultraviolet Photoemission Spectroscopy (UPS), X-ray Photoemission Spectroscopy (XPS), and X-ray Absorption Spectroscopy (XAS) at the nanoscopic level, thus leading to element selective imaging. Note that photoemission techniques require an additional analyzer enabling energy filtering of the photoemitted electrons. Information on the spatial distribution of the electronic structure, chemical composition and nature, or the local magnetization at the surface can be obtained. This technique is well suited to investigate modern nanostructures, fabricated using various methods (bottom-up or top-down, lithography, self-organized growth methods, etc.), at the nanoscopic level combining good spatial (few nm) and time (subnanosecond) resolutions. The increasing sophistication level of the nanostructures prompted considerable progress in high-resolution microscope probes that are sensitive to the compositional, electronic, chemical, and magnetic microstructure.

Grazing Incidence X-ray Scattering (GISAXS)

Geometry

In a typical GISAXS experiment the incident beam, of wave vector \mathbf{k}_i , makes a small incident angle α_i with respect to the sample surface. The incident beam is supposed to be monochromatic, parallel, and free from background contributions; the beam is cleaned using scattering slits whenever necessary. The direction of the scattered wave vector \mathbf{k}_f spans the (Q_y, Q_z) reciprocal plane (Fig. 1, top) that includes information about island size, height, and inter-island distance (or correlation). The pattern is observed on a 2D detector as far as possible from the sample in order to increase the reciprocal space resolution. Beam



Selected Synchrotron Radiation Techniques,
Fig. 1 (Top) Typical geometry of a GISAXS experiment. \mathbf{k}_i and \mathbf{k}_f are the incident and scattered wave vectors, respectively, yielding the momentum transfer (i.e., the reciprocal space vector) $\mathbf{Q} = \mathbf{k}_f - \mathbf{k}_i$. The angles α_i , α_f and $2\theta_f$ are related to the components of the momentum transfer, either parallel (Q_x and Q_y) or perpendicular (Q_z) to the sample surface, by the equations $\mathbf{k}_i = \mathbf{k}_0[\cos(\alpha_i), 0, -\sin(\alpha_i)]$, corresponding to the

wavelength $\lambda = 2\pi/k_0$ and $\mathbf{k}_f = \mathbf{k}_0[\cos(\alpha_f) \cdot \cos(2\theta_f), \cos(\alpha_f) \cdot \sin(2\theta_f), \sin(\alpha_f)]$ of equal modulus. \mathbf{k}_0 is known to be the incident wave vector modulus. (Bottom) (a) Fresnel transmission coefficients calculated at Cu K α photon energy for non-absorbing Cu, Ta and Cu using real δ and β values. (b) Penetration depth of X-rays calculated at Cu K α photon energy for Al and Ta. The incidence angle scale is normalized by the critical angle of total external reflection

stops hiding the direct and specular beam and rod are necessary because of the limited dynamical range of detectors (charge-coupled device, CCD camera in this particular case). The scattered intensity distribution depends on α_i , Q_y , and Q_z .

The experiment can be run in transmission for thinned samples or in grazing incidence conditions. Using focused X-ray beams and transmission geometries can provide a scanning

microscopy approach valuable to investigate inhomogeneities in the sample [9]. The grazing incidence geometry is better adapted for the study of nano-objects and their correlations without sample preparation.

Grazing Incidence

To understand the usefulness of grazing incidence some elements of photon-matter interaction have

to be recalled. Consider a sharp interface between vacuum and a surface of a material of wavelength-dependent index of refraction n .

The incident beam is supposed to be a linearly polarized plane wave that impinges on this surface at a shallow angle α_i with wave vector \mathbf{k}_i ; the reflected and transmitted beams leave at angles α_f and α_t with wave vectors \mathbf{k}_f and \mathbf{k}_t respectively. Snell's law writes as

$$\cos(\alpha_t) \cdot n = \cos(\alpha_i) \text{ and } \alpha_f = \alpha_i \quad (1)$$

As long as $n > 1$ total reflection cannot occur when light travels from the vacuum to the material, even if $\alpha_i = 0$. Fortunately, and unlike visible light, hard X-rays have an index of refraction less than unity and can be written as

$$n = 1 - \delta - i\beta \quad (2)$$

with $\delta = \frac{\lambda^2}{2\pi} r_e \rho_e$

$$\text{and } \beta = \frac{\lambda}{4\pi} \mu \quad (3)$$

in which λ is the X-ray wavelength, r_e the Bohr atomic radius, ρ_e the electronic density of the material, and μ the material linear absorption coefficient at the considered wavelength.

Consequently, for hard X-rays, the transmitted beam will be deflected always toward the sample internal surface, and total external reflection occurs on the vacuum side with a critical angle for total external reflection $\alpha_c \approx \sqrt{2\delta}$ in the 0.1 – 0.6° range because δ and β are, respectively, in the 10^{-5} and 10^{-6} ranges.

When $\alpha_i < \alpha_c$, the component of the transmitted wave vector normal to the surface becomes imaginary. The refracted wave is exponentially damped as a function of the distance below the surface and is an evanescent wave traveling parallel to the surface.

Considering the Fresnel transmission coefficient that writes as

$$T = |t|^2 = \left| \frac{2 \sin(\alpha_i)}{\sin(\alpha_i) - \sqrt{n^2 - \cos(\alpha_i)}} \right|^2$$

and the penetration depth given by

$$\Lambda(\alpha_i) = \frac{1}{-2k_0 \operatorname{Im}(\sqrt{\alpha_i^2 - \alpha_c^2 - 2i\beta})} \quad (4)$$

one can see that the surface sensitivity is strongly enhanced when α_i become close to α_c because the incident and reflected waves are nearly in phase. The penetration depth of the x-rays becomes nanometric, whatever the considered material, and the Fresnel transmission coefficient is maximal (see Fig. 1, bottom). One may note that higher absorption coefficients reduce slightly the Fresnel coefficient maximum and conversely decrease the penetration depth at any angle. Importantly, for $\alpha_i > 3 \times \alpha_c$ surface effects become negligible.

Grazing incidence scattering geometries provide simultaneously enhanced surface sensitivity and limit the unwanted bulk scattering (diffuse or not) that may otherwise overcome the surface or nanostructure feeble signal. It opens additional experimental possibilities as compared to simple specular reflectivity for which incidence and exit angles are kept equal. Grazing incidence is thus a configuration of choice in order to nondestructively investigate nanostructures deposited on – or buried under – a substrate.

Pattern Analysis

The interpretation and simulation of the measured scattering patterns is challenging and requires accurate models and, if size/distance distributions are considered, is demanding tedious computing. However, a complete modeling is not always necessary, and the data evaluation can go from quite simple for a crude first approximation to tedious when the full pattern reproduction is sought [10]. A practical illustration of all important terms is given in Fig. 1. In an extremely crude and semiquantitative approach, and neglecting particle shape, correlation, and anisotropy, one can deduce from the position and shape (width) of the diffusion lobes directly an approximation of the interisland distance (D), height (h), and size (R_0) as illustrated in Fig. 1.

The scattering signal contains a coherent contribution and an incoherent scattering term as soon

as the size distribution of the nanostructures is not monodisperse. The coherent term is the product of the Fourier transform of the nanostructure shape with the interference function $S(Q_{\parallel})$. For the incoherent scattering the two limit cases are (i) the Decoupling Approximation (DA), assuming no nanostructure correlations, and (ii) the Local Monodisperse Approximation (LMA), assuming full correlation between nanostructure sizes at a scale corresponding to the photon coherence length.

Hard X-ray photons interact weakly with matter, and except for perfect crystals or at glancing angles, the kinematical Born approximation (BA), which neglects multiple scattering effects, is valid. At angles below $3 \times \alpha_c$ dynamical effects of reflection and refraction at interfaces become important and have to be taken into account using the popular distorted wave Born approximation (DWBA) [11].

Within surface physics and nanoscience two approaches have to be distinguished depending on the level of interpretation that is sought:

- (i) The more intuitive effective layer Born approximation (ELBA), which applies for isotropic buried islands with simple shapes. The index of refraction is that of the effective layer in which the islands are supposed to be buried. In ELBA the resulting scattering pattern is the weighted sum of the scattering arising from each subelement since an assembly of objects can be considered as the sum of the individual scatterings as far as the relative intensities remain linked to the total number of electrons in each object within the x-ray coherence length.
- (ii) The DWBA, which applies in the vicinity of critical scattering geometries for islands deposited on a substrate. In the complete treatment, an anisotropic signal originating from island facets can also be considered. The effective form factor corresponds to the coherent interference of four waves corresponding to the four possible scattering events (weighted by the corresponding reflection coefficients) experienced by the incoming and exiting beams on a given island.

Within the ELBA, to first order and a rapid interpretation strategy assuming weak reflectivity, the scattered intensity $I(Q)$ is proportional to the product of three terms: the form factor $P(Q)$ which is the Fourier transform of the island volume, the interference function $S(Q)$, and the transmission factor $T(\alpha_f)$ that is supposed to simulate the dependence of the scattered intensity as function of α_f .

The interference function $S(Q)$ is related to the correlation length, which characterizes a distribution of islands on the surface. It is also related to the real space correlation function $g(r)$ via the formula

$$S(Q) = 1 + 2\pi \cdot \rho_s \cdot \int_0^r (g(r) - 1) \cdot J_0 \cdot (Q_c r) \cdot r \cdot dr \quad (5)$$

where ρ_s is the particle density per surface unit and J_0 the Bessel function of zero order. Many analytical different pair correlation functions can be used within models. One may consider, for example, a Gaussian pair correlation function (illustrated in inset of Fig. 1, top) given by

$$g(r) = \begin{cases} 0, & \text{for } 0 \leq r \leq R_0 \\ \frac{e^{-\frac{(r-D)^2}{\omega^2}} - e^{-\frac{(R_0-D)^2}{\omega^2}}}{e^{-\frac{(D_1-D)^2}{\omega^2}} - e^{-\frac{(R_0-D)^2}{\omega^2}}}, & \text{for } R_0 \leq r \leq D_1 \\ 1, & \text{for } D_1 \leq r \leq \infty \end{cases} \quad (6)$$

The most reliable input remains the direct evaluation of $g(r)$ from real space measurements made, for example, by complementary microscopy techniques.

Within the DWBA the effective form factor corresponds to the coherent interference of four waves corresponding to the four possible scattering events (weighted by the corresponding reflection coefficients) experienced by the incoming and exiting beams on a given island. The DWBA scattering intensity I_{DW} from islands deposited onto a substrate may be summarized as

$$I_{DW} = \left\{ \frac{|1 - n_\Delta^2|}{2 \cdot \text{Re}(1 - n_\Delta)} \cdot \left[\tilde{\chi}_\Delta(q_{//}, q_z) + R^f \tilde{\chi}_\Delta(q_{//}, -p_z) + R^i \tilde{\chi}_\Delta(q_{//}, p_z) + R^i R^f \tilde{\chi}_\Delta(q_{//}, -q_z) \right] \right\}^2 \quad (7)$$

where n_Δ is the index of refraction of the islands, $\tilde{\chi}_\Delta(q_{//}, q_z)$ is the form factor of the islands, $q_{//} = (q_x, q_y)$ is the component of the momentum transfer q parallel to the surface, q_z corresponds to the net momentum transfer perpendicular to the substrate surface, $p_z = k_z^f + k_z^i$ where k_z^f and k_z^i are the normal to the substrate plane wave vectors of the incoming and exiting waves. The reflectivity coefficients for the incoming and outgoing waves are denoted R^i and R^f , respectively. The Born approximation is restored with $R^i = R^f = 0$.

The specular reflectance of discontinued multilayer systems with interfacial roughness can be calculated by use of a recursive classical exact layer-by-layer method, based on Fresnel coefficients within the Parratt optical formalism and the dispersion model.

Interestingly, variations of electronic densities are the only requirement mandatory to enable the observation of GISAXS patterns. Objects (of lower electronic densities) in a larger electron density matrix will yield such signals. The extreme case will then be the presence of structured voids or pores. For reactive interfaces like NiO/Cu(111) [12], self-organized holes in the surface or the presence of Ni decorated corrals lead to GISAXS patterns similar to growing nanoparticles.

Size distributions can be included in the pattern evaluation in any framework. The obvious effect of the size distributions is to smooth the scattering pattern. Height distribution and cross-correlation between lateral size and height distributions are very difficult to extract from the GISAXS pattern.

Examples of Application

In 1989, GISAXS was introduced to study the dewetting of gold deposited on a glass surface and later, using synchrotron light metal agglomerates on surfaces and in buried interfaces were considered.

Determining the morphology of islands on a substrate, embedded clusters in matrices as well as the fabrication control of nanometer-sized objects can be successfully addressed using GISAXS. It has become a popular technique because it can be applied to a broad range of samples in various samples environments: investigation of samples in standard conditions [5, 10], porous layers, precipitates and quantum dots [13], growing particles in ultra-high vacuum conditions [4, 10], complex nanostructures like carbon nanotubes [14] and shape modifications upon various treatments. In recent years GISAXS studies have shifted towards the study of sample processing conditions and in-situ treatments [15–17]. The method in itself does not require any particular sample preparation – it is the thin film growth method requiring ultra-high vacuum environment. A major step has been realized in the last decade because of an in-depth understanding of the underlying phenomena and the availability of algorithms enabling the simulation of the experimental patterns. In the most advanced studies coherent GISAXS experiments are considered enabling full particle shape reconstruction [13, 18]. In general, GISAXS can be applied to characterize self-assembly and self-organization at the nanoscale in thin films.

The technique becomes highly interesting in situations where charge build-up is a problem like for investigating metal/insulating oxide systems. Electron based techniques are generally hampered by the insulating character of the sample and the charge build-up. Many such systems are characterized by Volmer-Weber or Stranski-Krastanov 3D growth of particles following a nucleation, growth and coalescence scheme. In GISAXS such a growth sequence leads to diffusion lobes moving towards the reflected beam in reciprocal space with respect to film thickness as illustrated in Fig. 2 for the Co/NiO(111) system. Even without pattern simulation the position and

shape changes of the diffusion lobes indicates directly the growth mode.

From sequences of images taken at various film thicknesses and sample conditions (like substrate temperatures or quality) one can extract the overall behavior of a metal/oxide interface and evaluate for example the effect of defects in the cluster and/or on the substrate. The study of Co deposits, performed in ultra-high vacuum conditions using substrates of different qualities, i.e., with variable amounts of nucleation centers is reported in Fig. 2. The experiment was performed on BM32 beamline (ESRF, Grenoble, France) using a beam energy of 10 keV. GISAXS was also combined with grazing incidence diffraction measurements (not shown here) to access the crystallinity of the sample. When the growth is performed at low temperature ($T = 450$ K) the Co clusters have numerous defects and poor crystalline quality. For Co thicknesses below 0.8 nm the islands diameters are similar whatever the surface quality or Co island crystalline quality. Thus in the nucleation regime, for cluster sizes below the onset of defects in the clusters, the number of nucleation centers on the surface does not lead to noticeable effects. Above 0.8 nm the diameters fairly diverge indicating that when coalescence starts to play an important role, the island mobility increases with temperature. This effect is consistent with the inter-island distance behavior of the islands. From the island height behavior it appears that the deposition at high temperature (good Co crystalline quality) on a substrate with numerous nucleation centers is equivalent to low temperature deposition (poor Co crystalline quality) on a high quality substrate. Such investigations can hardly be made using other techniques.

GISAXS is not only highly efficient to investigate *in situ* grown particles, it can also tackle very complex *ex situ* samples like carbon nanotubes (CNT) [14] organized in aligned long tubes perpendicular to the sample surface. Each tube includes on its top a Co nucleation particles used to promote the growth of the CNT. Figure 3 reports measurements made on beamline ID01 at the ESRF (Grenoble, France). The experimental pattern includes GISAXS scattering and reflectivity contribution. The calculated patterns required

the use of a core-shell C-Co structure inspired from transmission electron microscopy additional measurements. The height/size distribution where fairly large and the DWBA adequate parameters were retrieved from a first ELBA approximation. Such an analysis enables the evaluation of the effects of preparation conditions changes like the ammonia content in the reactive gas mixture [14].

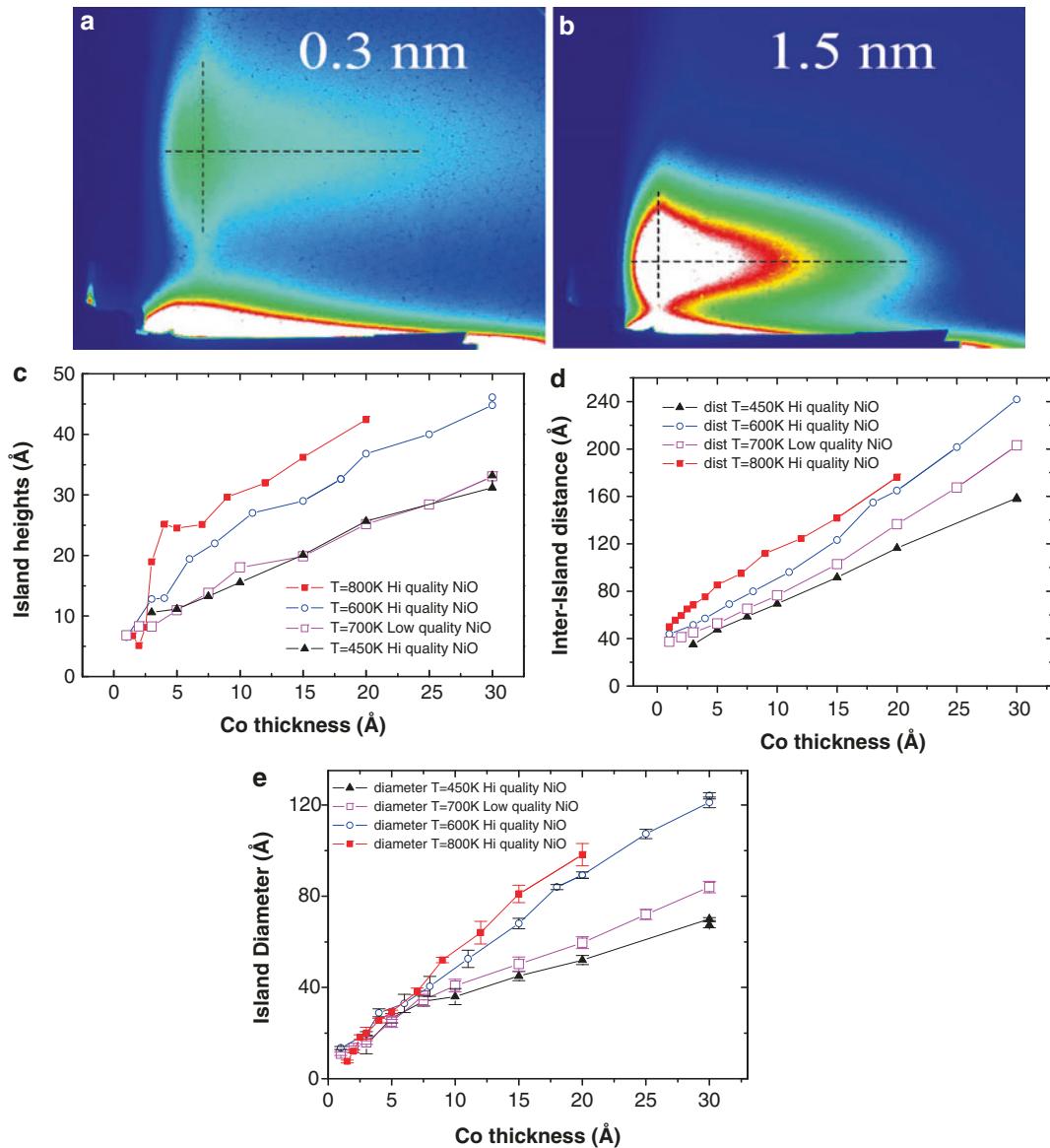
Recent GISAXS studies tackle a number of important issues concerning the driving forces of self-organization processes in block copolymers [19], nanoparticles [20] and organic electronics [17, 21]. The technique provides new and original contribution to these ongoing fields.

High Resolution Micro-Diffraction (μ -XRD)

In the local probe diffraction approach (μ XRD), it is necessary to illuminate with the X-ray beam only a single object. The probe size yields the lateral resolution. As a consequence, the volume of nanostructures probed by the X-rays will be highly reduced (from several 10^4 – 10^6 objects in a ‘standard’ XRD experiment to a single one), thus the detected scattered signal is reduced. Indeed, this one is proportional to the incident X-ray photon flux and the probed volume. In order to compensate for the very small probed volume, the X-ray beam has to be focused to small sizes. Several focusing alternatives are nowadays available.

A Short Introduction to Hard X-ray Focusing Optics

The following discussion is centered mostly around hard x-ray (several keV energy) optics used at synchrotron sources, although laboratory sources with X-ray beams of several 10 μ m are available and can be used for certain XRD experiments with local (lateral) resolution. But a stable, highly brilliant and low divergent X-ray source is a “must have” for obtaining very small and intense X-ray spots. The synchrotron x-ray source is generally situated far from the experimental station (several 10 m) which allows for large demagnification factors for the focusing optics.



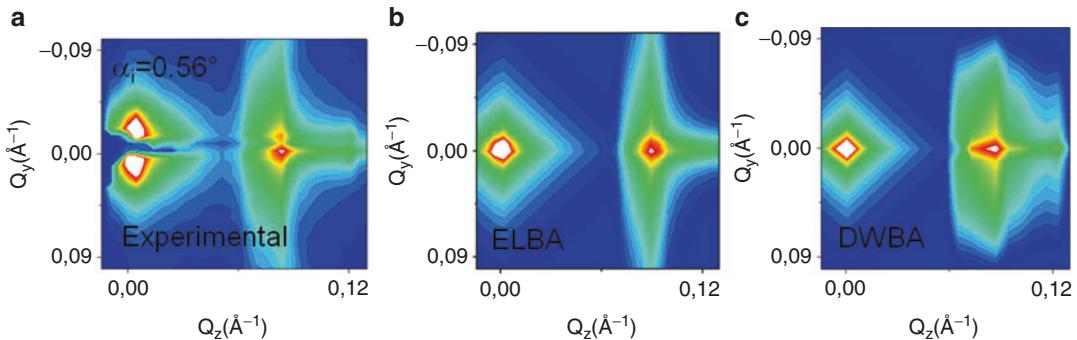
Selected Synchrotron Radiation Techniques,
Fig. 2 (a, b) Part of the GISAXS signal recorded on a CCD camera for (a) 0.3 nm and (b) 1.5 nm nominal thickness of Co deposited on NiO(111). The beam energy was 10 keV; the vertical of the CCD is parallel to the sample surface and the horizontal direction is perpendicular to it. The lead well at the bottom of the pictures hides the direct and reflected beams. The dashed lines are guide for

the eye to evidence the changes in shape and position of the GISAXS signal. (c, d, e) Morphology of Co islands grown on NiO(111): Island height (c), inter-island distance (d) and island diameter (e) with respect to the deposited Co thickness, as extracted from GISAXS measurements. Hi- and low-quality NiO substrates had mosaic spreads of 0.05° and 0.5° respectively

Consequently, major progress was done at synchrotron facilities, where these focusing devices can be used up to their limits: beam sizes of several 10 nm have been demonstrated and highly

intense beams in the sub- μm size range are available at a number of synchrotrons [1].

A detailed description of the various focusing/collimation schemes for X-rays proposed in the



Selected Synchrotron Radiation Techniques

Fig. 3 (a) Experimental GISAXS pattern of CNTs prepared with 1 % ammonia concentration in the reactive gas mixture, collected at $\alpha_i = 0.56^\circ$. The sample surface is vertical and located on the *left side* of the pattern. A horizontal beam-stop stops the intense direct and specular beams. (b) ELBA pattern including Co cylinders,

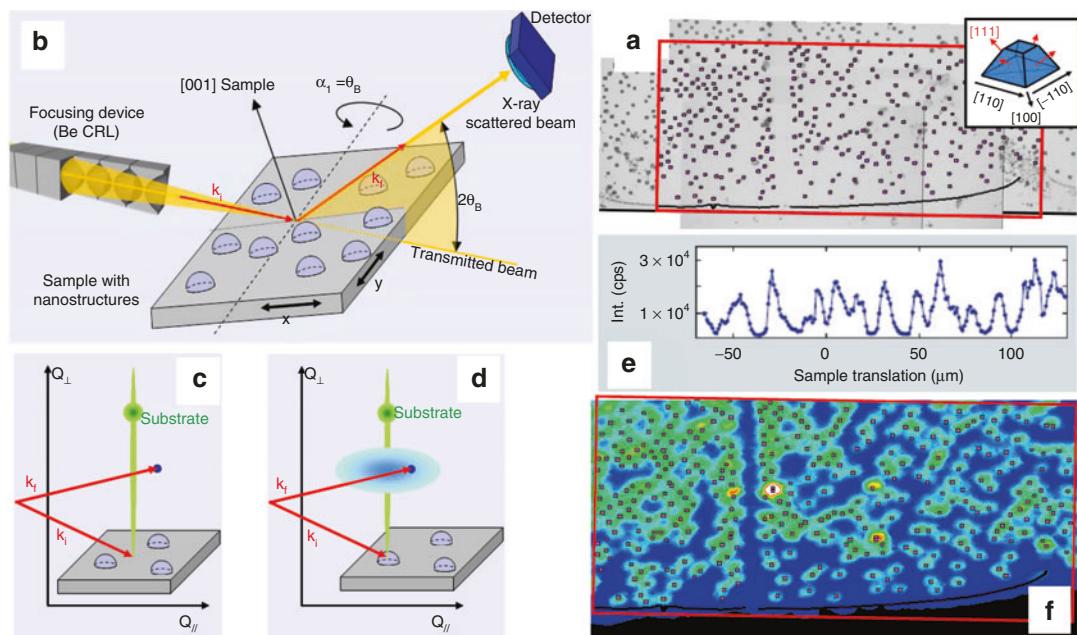
C tubes, Co-C intermixing (Co-C core-shell 14.8 nm average object and intermixed C-Co intermixed tubes); Co and C radii 2.5 and 3.3 nm respectively, inter-tube distance 55.6 nm, (c) DWBA pattern using the parameters retrieved from the ELBA model. The colour scale is logarithmic and identical for all patterns

last years can be found in Ref. [6]. Their respective advantages and drawbacks have to be carefully considered when designing a micro/nano X-ray beam experiment.

μ -XRD Setup and Experimental Approach

Using X-ray focusing optics, it is possible to build various X-ray microscopes working with different contrast mechanisms, or combinations of them: fluorescence, luminescence, transmission, diffraction. The particular case in which a diffraction contrast is used will be briefly described here. The setup is a combination of a X-ray focusing element and an accurate diffractometer. It consists of a scanner stage used for laterally positioning the sample in the X-ray beam, coupled to precise rotation stages both for the sample angles (incidence, azimuth, tilts) and the X-ray detector (at or close to Bragg diffraction angle) for probing, in the reciprocal space, any desired lattice parameter. Optionally a high resolution optical microscope pointing at the precise position of the diffractometer's centre (and X-ray beam) is used for sample pre-aligning and finding of the region of interest. The instrument can function in two ways: one in which the sample surface is *imaged point by point like in a scanning probe technique*, and a second one in which, for particular regions (microstructures) of the sample,

detailed *high resolution μ -XRD* data are recorded. Both approaches will be illustrated hereafter. Figure 4 shows the principle of the method. The sample considered here consists of SiGe islands (square based truncated pyramids, inset Fig. 4, panel a) [22]. Panel (a) shows an optical image of the sample surface, image on which the individual SiGe islands can easily be identified (dark squares). The x-ray beam is focused onto a small spot on the sample. The elastically scattered x-ray photons are collected by the detector. By rotating the sample and detector angles, the intensity distribution around different reciprocal lattice point positions can be probed [4]. This is done by appropriately setting the Bragg angles (panel (b)), beam impinging incidence angle given by the \mathbf{k}_i vector and the position of the detector given by \mathbf{k}_f . If the sample is laterally translated in the X-ray beam without changing any of the angles, one expects to see differences in the scattered intensity depending if an island or the bare substrate is illuminated. In the case of the substrate, the intensity distribution shows a sharp Bragg peak (panel (c)), and a streak perpendicular to the sample surface, the so called crystal truncation rod (CTR) [4, 5]. If the focused x-ray beam hits an island, additional and broader features will appear in reciprocal space at different position (panel (d)), since a different lattice parameter, the one



Selected Synchrotron Radiation Techniques,
Fig. 4 Principle of the SXDM approach (see also Refs. [6, 23, 24]). (a) Optical image of the SiGe/Si pyramids sample. The inset shows a sketch of the shape of the SiGe square-based truncated pyramids. The red rectangle shows the area imaged using the μ XRD approach (panel f). The optical image, obtained in back-reflected light, makes that the side facets are deflecting the light out of the optical axis of the objective, thus appearing dark. Each dark square represents a single SiGe island. (b) Illustration of the setup, k_i and k_f being the scattering vectors. (c, d) The x-ray spot (represented by the tip of the k_i arrow) illuminates the substrate only or a single island (respectively). Only in the latter case a broad x-ray diffraction signal (different

lattice spacing distribution inside the island) is observed, at a different position with respect to the substrate signal. Tuning the diffraction angle such to be sensitive in the reciprocal space to a position characteristic to the islands (*dot* pointed by the tip of the *arrow kf*) and scanning the sample laterally, higher intensity is observed only when the X-ray spot illuminates an island (panel e). If this is done in both x and y directions, a map of the surface of the sample (SXDM) is obtained (panel f), with enhanced intensity (*green* and *yellow* color) when an island is probed. Overlaid to this image, the position of the SiGe islands extracted from the optical image (shown in panel a) is superposed as red squares

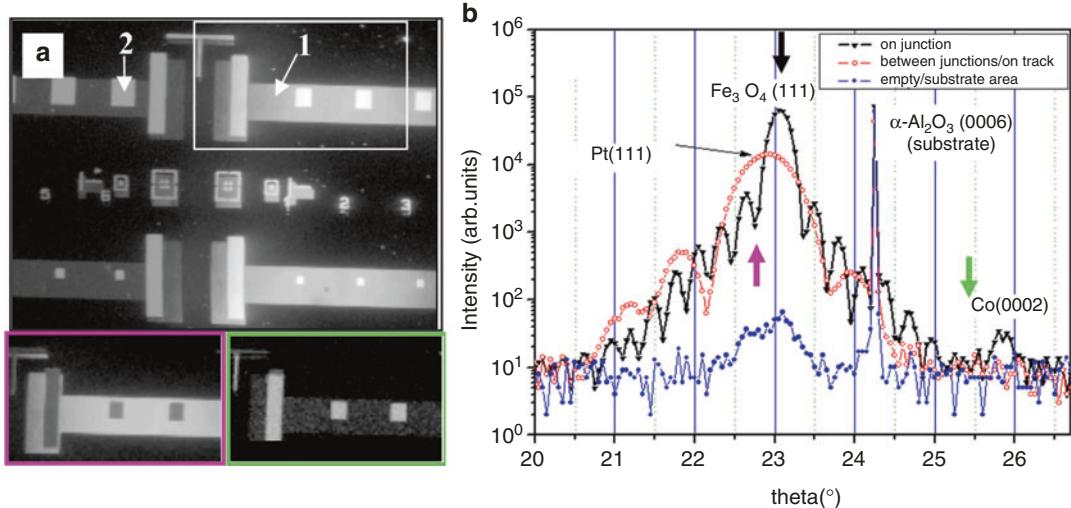
of the SiGe island, is probed. The origin of this different lattice parameter can be multiple, some of the possible causes can be: a various composition, strain, shape, relaxation, etc.

Examples of Applications

Scanning X-ray Diffraction Microscopy (SXDM): Imaging with Diffraction Contrast

It is now straightforward to perform a “microscopy”-like experiment, the result being an image of the surface of the sample. The diffraction angles are fixed at the expected values to record diffraction from the crystalline planes

inside the nanostructures and the position of the sample is scanned laterally while recording, at each point, the scattered intensity (Fig. 4 panel e)). The resulting contrast is diffraction based and is related to the islands presence. If the mapping is performed in both directions (x and y), the resulting map is the result of the Scanning X-ray Diffraction Microscopy (SXDM) approach. Panel (f) shows a SXDM image (color scale) superimposed to the optical image of the same region of the sample (from panel (a)). It is possible to find precisely a particular object: in this case, the region close to the ‘defect’ exhibited as the vertical zone without any SiGe islands. Markers could also be envisaged. Indeed, it is of utmost



Selected Synchrotron Radiation Techniques,
Fig. 5 (a) Scanning μ XRD image (SXDM) of the MTJ surface showing defects due to mask align in the optical lithography process. Areas where only the Pt layer (1) remained after the lithography process or the full MTJ (2) is present are shown. (b) μ XRD data (θ - 2θ geometry) at different locations on the sample: \blacktriangledown on a MTJ; \circ on

contact track (Pt buffer); \bullet on the empty area (substrate). The colored arrows point to reciprocal space positions characteristic of the different crystalline structures (Pt, Fe_3O_4 and Co), positions for which SXDM images of the sample surface (shown in panel a) were performed (see also the colored frame insets in panel a)

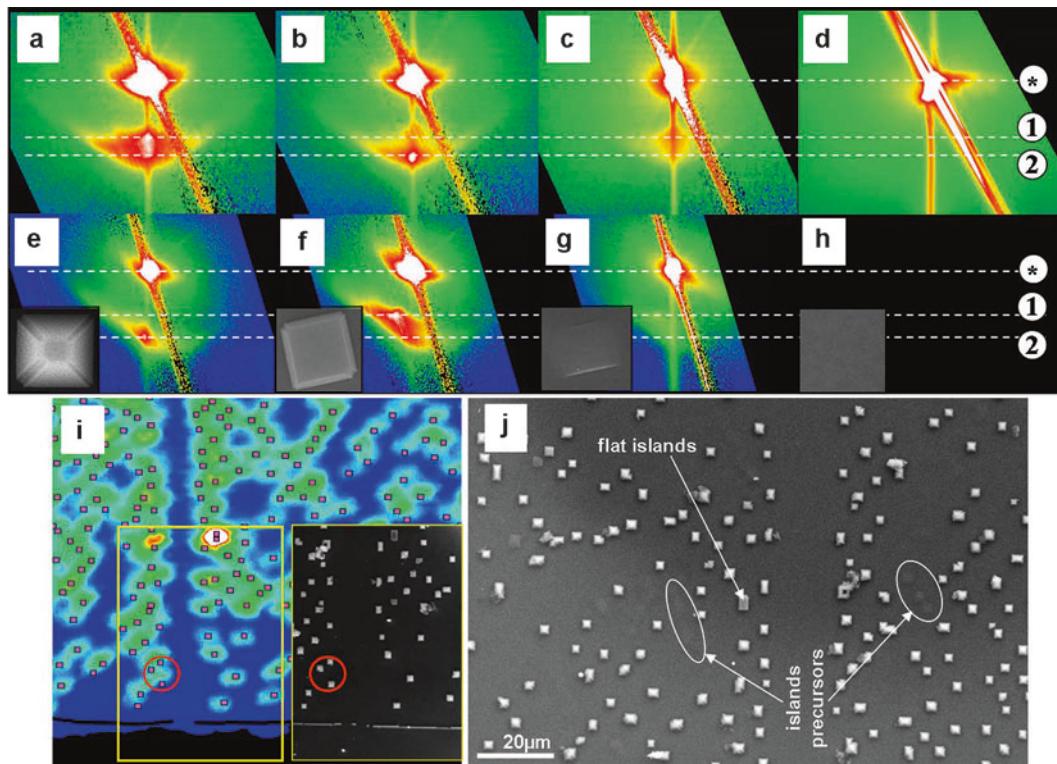
importance not only to be able to measure an isolated object out of an ensemble, but to precisely know which particular object was measured: the properties of these SiGe islands can change dramatically from object to object and can be correlated with their shape obtained by \blacktriangleright SEM.

The second example below is showing the power of the method and its sensitivity not only to (local) crystalline changes, but to thickness as well. The sample (metal oxide Magnetic Tunnel Junction, MTJ) consists of a stacking of several metallic and oxide layers on which structures of lateral size in the 10–100 μm range were made by optical lithography [25]. This particular sample suffered a mask misalign during one of the lithography steps. Although this could be easily detected by simple optical microscopy examination of the surface, μ XRD can also bring some information about how this happened and which layers suffered; this sample is a good example to show how the different structures can be imaged and differentiated. The corresponding XRD recorded at these different locations are shown in Fig. 5 panel (b). Signal originating from

the various crystalline structures of the layers (Pt, Fe_3O_4 and Co) can easily be differentiated and used as probe in imaging the surface of the sample, after tuning the diffraction angles to these positions (as highlighted by the colored arrows in panel b). The contrast on the surface images changes accordingly. Moreover, sensitivity to thickness is achieved: the arrow pointing to the maximum of the Pt signal also corresponds to a minimum contrast (interference) for the Fe_3O_4 layer. This explains the reversed contrast obtained in the left bottom panel (a) of Fig. 5. For more details, the reader can see Ref. [26] and references therein.

High Resolution X-ray Micro-Diffraction (HR – μ XRD) and Combination with SEM

With the approach depicted above it is possible to measure not only an *individual* object but a very *particular* one [6, 23, 24]. The case of the SiGe islands will be used as example. On the SXDM image of the surface of the sample (Fig. 4), individual objects are chosen and measured in HR- μ XRD. Once the lateral positions of the



Selected Synchrotron Radiation Techniques,
Fig. 6 (a–h) Reciprocal space maps (logarithmic intensity scale) in the vicinity of the (004) and (115) Bragg peaks, recorded using a X-ray focused beam. The representative SEM image for each type of object is shown in the inset. The Si substrate Bragg peak is indicated by (*). (i) Image of

sample are fixed such that the X-ray spot illuminates the object of interest, the angles (sample and detector) are scanned in order to describe the reciprocal space in the vicinity of the chosen Bragg position. The recorded data show characteristics allowing to be grouped in three categories corresponding to three types of SiGe objects: the scattered intensity might vary slightly for the same type of object (depends on the particular imaged object), but, for one family the same characteristics are found. The differences are from one type to another. Figure 6a–h shows such measurements. The differences from the three mentioned types are visible both close to the (004) and (115) Bragg peaks. Since these μ XRD measurements were performed at known positions of the sample (cf. image of the surface, as described in Fig. 4), it is possible, at the end of the diffraction

the sample surface, performed with diffraction contrast (SXDM), compared to a SEM image of the same area, with particular objects identified (red circle). (j) SEM image (details) of the sample, around the region with the ‘vertical path’ free of SiGe islands, in the middle of the image

experiment, to image in detail precisely the objects in question. The corresponding SEM images of the objects are reported as insets in Fig. 6a–h. Comparing the scattered intensities measured for SiGe pyramids and flat islands, some major differences can be highlighted: the intensity distribution close to the SiGe peak is changing. If the truncated pyramid exhibits scattered intensity concentrated around two positions (two maxima, labeled (1) and (2) in the figure), the flat islands show distributions mostly around the position (2). In the case of the (004) map, the position (1), closer to the Si Bragg location, shows the presence of a lattice parameter (out of the surface plane) closer to the Si one, i.e., indication of a lower Ge concentration. The signal at position (2) indicates a higher Ge concentration and a larger lattice parameter.

The availability of two types of maps, around symmetric (004) and asymmetric (115) Bragg peaks, bring complementary information: the peak's position on the (115) maps has also an in-the-surface plane component of the SiGe lattice parameter [24].

This example shows the need and the power of combining several methods on the very same micro/nano-object, for detailed characterization and modeling: the shape and the sizes are deduced from SEM images and then injected into a model used to simulate the diffraction data (finite elements methods for strain distribution, then semi-kinematical scattering theory), more details can be found in Refs. [6, 24].

In-situ Combination of μ XRD with Atomic Force Microscopy (AFM)

The previous example showed the combination of μ XRD with (ex-situ) ► SEM images, yielding to an understanding of the structure of microscopic semiconductor samples. The next example shows the in-situ combination of a μ XRD experiment with an Atomic Force Microscope (AFM) [27, 28]. This approach allows measuring in the same time and for the same individual small object the mechanical behavior, the internal strain and the response to external stress. Such measurements, especially when approaching the nanoscale, are of the highest importance for materials characterization, both in the elastic and the plastic deformation regimes.

The SiGe island sample described before was used as well as model sample to prove the principle of this experiment. The experimental setup is schematically shown in Fig. 7a: an ► AFM is mounted on the diffractometer and the focused X-ray beam is aligned with the apex of the tip. During point by point mapping of the sample surface with the X-ray spot (SXDM), the electron current induced in the AFM tip by the X-ray beam is also recorded. The resulting two contrast images (acquired simultaneously) are then combined (green and red colors respectively) to yield the image shown in panel (b). Panel (c) shows precisely the same sample area imaged in SEM (ex-situ) previous to the experiment – some of the missing islands in panel (b) were ‘destroyed’ by

plastic deformation attempts during the first part of the AFM experiment.

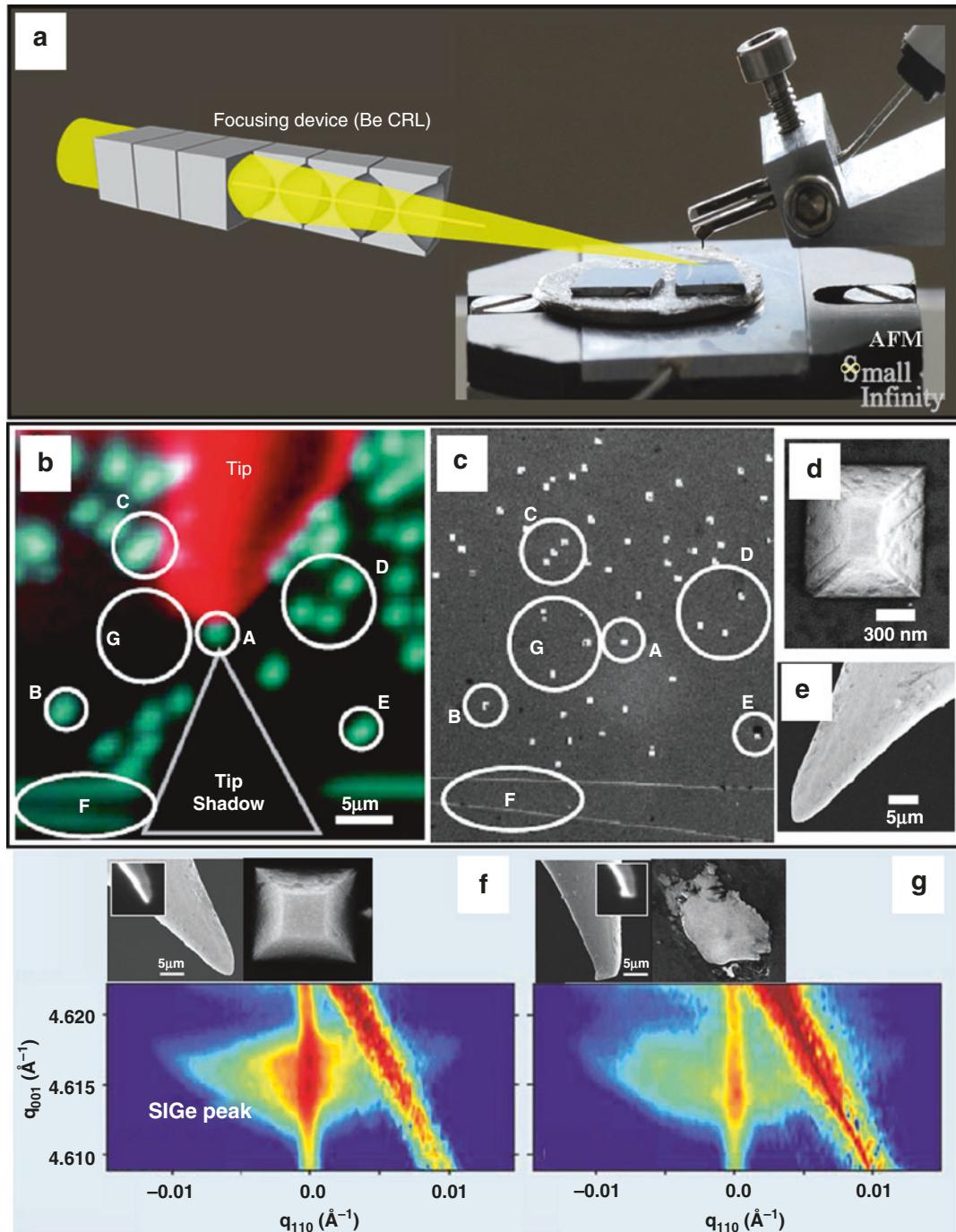
The applied force (and thus the pressure) can be extracted from the changes in the resonance frequency of the AFM tip and the lattice parameter (average value over the whole island) is extracted from the Bragg position in μ XRD experiments, at each applied external pressure. It was also possible to access the plastic regime. Figure 7 (panels f and g) shows μ XRD measurements (reciprocal space maps) close to the (004) Bragg position for a SiGe pyramid before and after pushing. The insets show the corresponding SEM images of the island and AFM tip before and after the experiment. Although, the AFM tip bends, the applied pressure can be large enough to completely destroy the SiGe island, which result in a major change of the μ XRD signal.

This example validates the use of the coupled μ XRD and AFM not only for studying the elastic properties of small structures, but also for in-situ indentation of materials (► Nanoindentation) (a stiffer tip, and with an adapted shape for indentation, should then be used). In such an indentation experiment, by applying laterally resolved μ XRD, the resulting strain field around the indented area could be mapped and modeled.

X-Ray PhotoEmission Electron Microscopy

XPEEM Instrumentation and Operating Principle

XPEEM microscopy is a parallel imaging method that combines X-ray electron spectroscopy and electron microscopy. It is based on soft X-ray-in/Electrons-out principle and was originally pioneered by B. Tonner. The sample is illuminated by a monochromatic soft X-ray beam (<1500 eV). The method does not require extreme focusing of the X-rays. Moderate focusing (few μm^2) using dedicated X-ray optics (Kirkpatrick – Baez, Volter mirrors, etc.) is used so that the beam spot matches the useful microscope field of view. After the absorption of the X-ray photons, the emitted photoelectrons are collected, magnified and projected onto a detector



Selected Synchrotron Radiation Techniques,
Fig. 7 (a) Cartoon of the combination of μ XRD with AFM (simplified setup). Once the X-ray focused beam aligned with the tip of the AFM, it is used to ‘image’ (high resolution XRD) the object placed beneath the tip. (b) superposed maps (images) of the sample, with the X-ray beam aligned on the AFM tip and a single SiGe

island placed just beneath the tip: in green the detector signal (sensitive to SiGe) is recorded, while in red is the electron current induced in the tip when this one is hit by the X-ray beam. (c) SEM image of the sample showing precisely the region measured in panel (b). The different areas on the sample are labeled. Note the absence of the islands labeled ‘G’ – during the

using a dedicated (electrostatic or magnetic) low electron energy microscopy column.

There is a certain number of XPEEM microscopes available commercially. A version that combines PEEM and LEEM microscopy is nowadays the most popular one used with synchrotron radiations sources. This apparatus is very powerful because it allows to combine in a single instrument structural (LEEM) and spectroscopic (PEEM) methods, enabling a true multi technique approach to study surfaces, buried interfaces and nanostructures.

The electron column used in the XPEEM microscopy is similar to those used in Electron microscopy, the main difference arising from the use of low kinetic energy photoelectrons. The XPEEM instrument in its simpler version consists of an objective lens, a projective lenses, an image detector and a contrast aperture (see Fig. 8). A first image, collected and magnified by the objective lens, is formed in the back-focal plane of the objective lens. An aperture placed within this plane (contrast aperture) allows to reduce the energy spread of the photoelectrons and limits their angular dispersion. The image is further magnified by a set of projective lenses. The final image is collected via a detector consisting of a Micro-Channel plate, a fluorescence screen and a CCD camera. In the most sophisticated instruments, a certain number of lenses are additionally implemented: corrector lenses for astigmatism, retarding grid as high-pass energy filter, field lenses, etc.

Although the incorporation of an energy filter in XPEEM microscope is not mandatory for imaging, its installation in the most recent and advanced microscopes opens up the way for core

level and valence band photoemission microscopy. The energy filter is not only helpful for imaging with primary photoelectrons but also useful for secondary electron imaging. It allows to select a narrow energy window around the maximum of the secondary electron energy distribution and, thus, to improve the spatial resolution without unacceptable loss of transmission.

The main requirements for X-ray microscopy, in particular for PEEM/LEEM, are the spatial resolution and the contrast mechanism.

Spatial Resolution

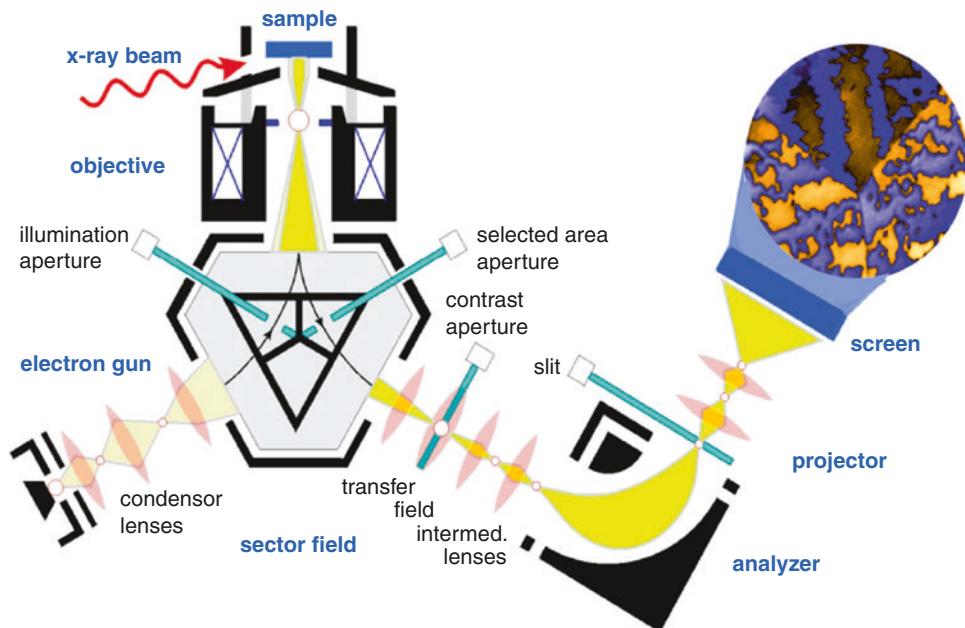
A spatial resolution of 22 nm using synchrotron radiation has already been achieved and a resolution of 6 nm using LEEM. A key limitation to the performance of electron-optical X-ray microscopes is the severe chromatic and spherical aberration of the immersion lens used. This problem constrains the X-PEEM microscope to a fairly small transmission in an attempt to offset these problems. Rempfer's group has demonstrated a solution to this problem, by designing and testing an electron mirror, which has spherical and chromatic aberrations of similar magnitude to the objective lens, but opposite in sign. When used in combination with a properly designed objective lens and a highly symmetric beam separator, the aberration can be highly reduced. Therefore both transmission and resolution can be increased. Several projects using correcting mirror are under development [29].

Contrast Mechanism

The contrast mechanism used in XPEEM microscopy arises from the spectroscopic capabilities and the characteristic binding energies of the

Shark Skin Effect, Fig. 7 (continued) experiment and previously taking this image, these islands were ‘smashed’ (destroyed) or laterally displaced using the AFM tip. (e) SEM image of a SiGe single pyramid of $\sim 1 \mu\text{m}$ lateral size. (e) Image of the blunt tip used for pushing experiments. Note the relatively large apex radius of curvature (with respect to standard AFM tips) – this feature makes easier pushing on the small objects. (f) Reciprocal space map (close to the (004) Bragg peak) for a single SiGe island.

The Si substrate Bragg position is not shown on the map ($q_{\text{Si}(004)} = 0.4628 \text{ nm}^{-1}$). (g) Same map after pushing with the AFM and reaching the plastic regime (destroying the SiGe island). The intensity distribution changed completely. The insets show the SiGe island and the AFM tip (SEM images) before and after plastic deformation. The electron yield images of the tip are also shown for each case



Selected Synchrotron Radiation Techniques, Fig. 8 Schematic sketch of the LEEM-PEEM microscope

atomic core electrons. The illumination of the specimen with X-rays excites a broad electron spectrum consisting of primary photoelectrons and inelastically scattered secondary photoelectrons. Two operation modes can be therefore performed:

- XPS (X-ray Photoelectron Spectroscopy). The XPEEM detects and selects, using an appropriate energy filter, the primary photoemitted electrons from atomic core levels. The photon energy is fixed and the kinetic electron energy is given by: $E_K = h\nu - E_b - \phi$, where $h\nu$ is the photon energy, E_b is the core level binding energy and ϕ the work function i.e., the energy barrier that the photoelectrons have to overcome to escape from the sample surface.
- XAS (X-ray Absorption Spectroscopy). In this case the XPEEM detects the secondary photoelectron as function of the impinging photon energy. When the photon energy matches the absorption threshold of the element, the photoelectron spectrum shows strong resonances due to electron transition arising from the core level to the unfilled valence band states.

The choice of XAS or XPS modes is mainly made on the basis of sample depth probe requirements ($1/e$) which are directly correlated to the electron inelastic mean free path (e) and the photoelectron kinetic energy. XPS allows to achieve a high surface sensitivity with a probing depth below 1 nm, meanwhile the XAS is more bulk sensitive (<10 nm).

In both cases, the XPS or XAS intensities are proportional to the number of emitter atoms within the probing depth, and thus provides direct and quantitative mapping of the chemical composition. Additional information can be also obtained from the lineshape analysis as they are a fingerprint of the emitter chemical state (valence state, site location, etc.). Finally, local spectroscopy can be performed selecting a small region of the specimen: local spectroscopy (μ -XAS or μ -XPS), photoelectron diffraction (μ -XPD), angle resolved photoelectron spectroscopy (μ -ARPES), etc.

The high spatial resolution of XPEEM microscopy, its chemical sensitivity and spectroscopic capability open up a broad field of application which extends from thin film studies,

nanostructures, magnetism to more exotic applications for surface science such as in tribology and geology.

Examples of Application: Magnetic Imaging

The interest in magnetic domain imaging in the nanometer range has been rapidly increasing during the last decade. A considerable impetus is coming from the development of high-density magnetic storage devices and from the forthcoming achievement of spin electronics. In order to tailor the magnetic behaviour of these systems to specific needs (for instance a certain response to magnetization reversal), a detailed understanding of the structure and of the dynamics of magnetic domains is mandatory. In addition, the thin film nature of such devices emphasizes the surface aspect of magnetism. This situation requires magnetic domain-imaging techniques which combine surface sensitivity and high spatial resolution. Moreover, for many applications element specificity is even more important than high lateral resolution. Magnetic storage media or building elements of spin-electronic devices are often composed of several chemical elements or intermetallic compounds, each of which distinctly contributes to the magnetic behavior. All these requirements pose a considerable challenge to conventional magnetic domain imaging techniques such as magneto-optical Kerr microscopy, Lorentz microscopy, scanning electron microscopy (SEMPA), Magnetic Force Microscopy (MFM), etc.

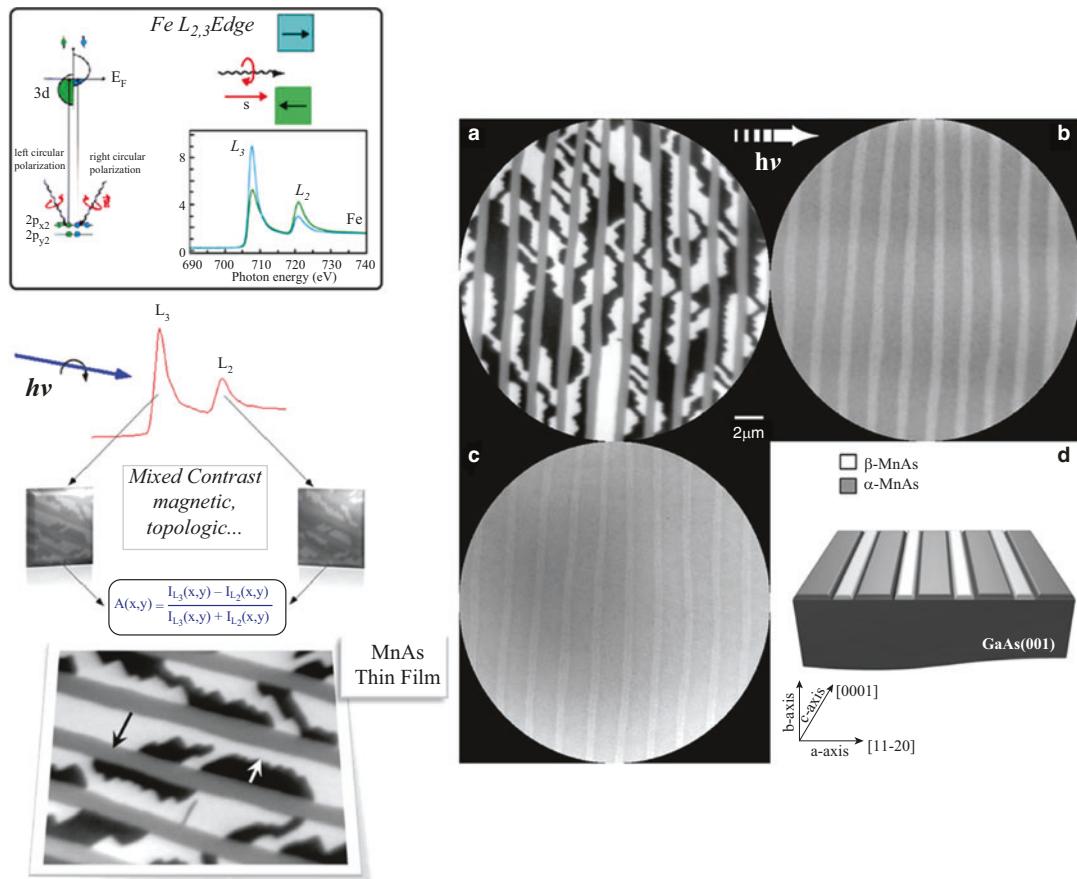
XPEEM magnetic microscopy is today a good candidate for an ideal surface magnetic imaging technique, as it combines the magnetic and element selectivity with a spatial resolution below the size of the magnetic domains. One may identify three important length scales for magnetic imaging which span over two orders of magnitude. The first one is about 1 μm , set by the size of lithographically manufactured magnetic cells such as in spin valve read heads or magnetic memory cells. The second one is about 10 nm, corresponding to the crystallographic grain size of

typical magnetic materials and to domain walls. The last one is 0.1 nm, i.e., the atomic size. The actual spatial resolution of the XPEEM microscope allow to image straightforwardly the magnetic domain, the further improvements of the resolution may allow access to the second characteristic length scale of 10 nm.

The elemental specificity in XPEEM magnetic microscopy arises from the characteristic binding energies of the atomic core electron, as mentioned above. Both X-ray photoelectron spectroscopy (XPS) and X-ray absorption (XAS) can be used for magnetic imaging, although the latter is the most used because it is less demanding in terms of instrumentation and photon flux.

The use of polarized synchrotron radiation enables studies of the electronic and magnetic anisotropies, and thus allows magnetic contrast for the XPEEM [30]. A simple description of the photon polarization by a biaxial vector for linear polarization and a vector for left/right handed circular polarization is the physical basis for probing various anisotropies of the sample. In general, linearly polarized light can only detect anisotropy of electronic charge. In contrast, helicity resolved circularly polarized light can measure a dipolar or vector quantity, in the present case the modulus and direction of the electron angular moment and spin.

For magnetic XPEEM spectromicroscopy, X-ray Magnetic Circular Dichroism (XMCD) is exploited. This effect is widely used to determine the size, the direction and the anisotropy of the atomic magnetic moments in magnetic material at a macroscopic scale. In the simple figure of a 3d transition metal, the magnetisation or more precisely the spin moment, is given by the imbalance between the spin-up and spin-down electrons (or holes) within the d shell and bellow the Fermi level. The use of circularly polarised light will make the absorption process spin-dependant and therefore enable access to the spin moment. In other word, the spin-split valence band will act as a detector for the spin of the excited photoelectron: the XAS intensity is simply proportional to the number of empty d states of a given spin. The measurements of the XAS signal anisotropy with respect to the direction of the light for a given



Selected Synchrotron Radiation Techniques,

Fig. 9 (Left panel) *Top*: Principles of X-ray magnetic circular dichroism (XMCD) in the case of L Fe edge. The XMCD spectrum reflects the anisotropy of L absorption edge with respect to the relative direction between the incident light and the magnetization direction at a given circular polarized light. Bottom: Principles of XMCD-PEEM in the case of a MnAs thin film. The measurements are performed at the Mn L edge (see text for explanation). (Right) XMCD and XMLD-XPEEM images. MnAs

sample thickness 330 nm: (a) XMCD image at the L_3 Mn edge highlighting the FM Mn domains in the α -phase. (b) XMLD image taken using the L_3 multiplets asymmetry at vertical linear polarization evidencing the AF domains. The gray level corresponds to the isotropic α -phase and the light region to the orthorhombic β -phase. (c) LEEM image taken at the same sample area showing the α - β phase coexistence. (d) Schematic sketch of the FM/AF configuration in the MnAs thin films

circular polarisation enable directly the access to the projection of the magnetisation as illustrated in the Fig. 9. In XMCD-PEEM, a set of two images are acquired at respectively the L_3 and L_2 edges for a given light helicity and direction. These two images will contain a mixture of different contrast: magnetic, structural, chemical.... In order to cancel all the contrast mechanism except the magnetic one, an asymmetric image is calculated by subtracting the above two mentioned data. This

image reflects directly the local variation of the $L_{2,3}$ asymmetries and therefore gives access to the projection of the magnetisation moment with respect to the light direction: the black and white regions in the image reflect the domains where the magnetic axis is aligned parallel or anti parallel to the direction of the light at fixed polarization. The grey area corresponds to domains where the magnetic axis is perpendicular to the direction of the light or that have no magnetic moment. It is

worth noting that the same XMCD-PEEM image can be obtained by varying the circular light helicity and keeping the photon energy fixed at the same absorption edge (L_3 or L_2).

The study of antiferromagnetic (AF) surfaces and interfaces has posed an even larger challenge because conventional techniques are mainly bulk sensitive and antiferromagnet do not carry any net external magnetic dipole moment. These limitations were overcome by the use of X-ray Magnetic Linear Dichroism (XMLD) spectroscopy. In contrast to XMCD which directly measures the magnetic moment, XMLD measures the projected value of the square of the magnetic moment. XMLD can therefore be applied for all uniaxial magnetic system, i.e., antiferromagnets as well. Recently it was shown that XMLD spectroscopy in conjunction with X-PEEM microscopy is capable of imaging the detailed antiferromagnetic domain structure of a surface and interface [31]. This has been shown on a NiO(001) thin film and cleaved sample. The XMLD-PEEM images reveal antiferromagnetic contrast corresponding of the different in-plane projection of the antiferromagnetic axis.

The ability to combine in the same instrument XMCD and XMLD XPEEM microscopy with high spatial resolution and chemical selectivity has open a new route to investigate magnetic materials, as illustrated in the following examples.

Surface Magnetism in MnAs Thin Films

Ferromagnetic (FM) MnAs is a promising candidate for electrical spin injection into GaAs and Si based semiconductors, since it exhibits a large carrier spin polarization, small coercive field and relatively high saturation magnetization and Curie temperature. Bulk MnAs is FM at Room Temperature (RT) (α phase) and shows close to 40 °C a first order transition to the paramagnetic β phase. On the contrary, epitaxial MnAs films on GaAs substrate, which are more appropriate for the injection applications, show at RT the coexistence of both phases. The phase coexistence results in the formation of self-organized stripes of alternating α and β phases. The XMCD-PEEM image of Fig. 9 (right) illustrates this phase coexistence. The black/white regions correspond to

ferromagnetic domains in the α -phase, the vertical grey stripes correspond to the non-magnetic α -phase. The easy magnetisation direction is perpendicular to the stripes direction, pointing in opposite directions in the black and white regions. The period of the stripes is closely correlated to the film thickness ($4.8 \times$ thickness). The phase coexistence is due to anisotropic strain applied by the GaAs substrate. As a consequence the magnetization is pointing not along the stripe direction as expected from shape anisotropy considerations but is perpendicular to it, predominantly in-plane. This leads to a complex, thickness-dependent magnetic domain structure in the interior of the α -MnAs which is reflected in the complexity of magnetic images of the surface of the film.

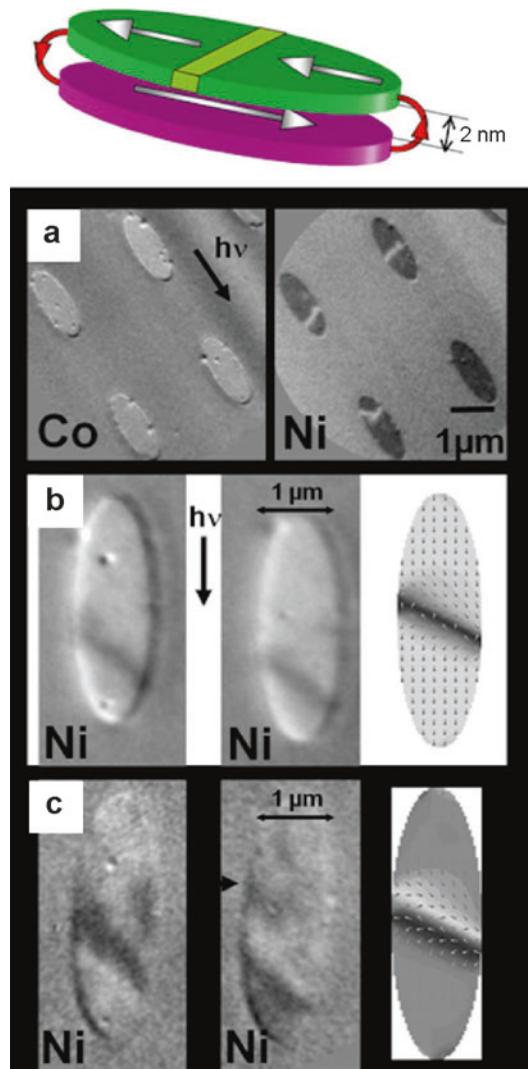
The nature of the non-magnetic β -phase is still controversial. In the coexistence range exchange bias and giant magnetoresistance effects of possible practical importance have been reported. Until recently β -MnAs has generally been considered to be paramagnetic but the effects just mentioned and first principles calculations suggest antiferromagnetism. A more thorough first principles calculation comes to the conclusion that β -MnAs is paramagnetic. The XMLD-PEEM microscopy enables to elucidate this controversy. Figure 9 evidences a weak antiferromagnetism in the β -phase which explain the reported exchange bias and giant magnetoresistance [32].

Micromagnetism of Patterned Nanostructures and Magnetic Tunnel Junction

Magnetic tunnel junctions (MTJ) are probably one of the most studied devices in the so called spintronics research field. They are usually composed by two active ferromagnetic electrodes separated by a thin insulating layer (<3 nm) that act as a tunnel barrier for the spin polarised electrons i.e., a spin dependant tunnel current exists and can be induced between the two ferromagnetic electrodes. The most attractive property of MTJs concerns their strong electrical resistance variation with the magnetic configuration of the electrodes, the so called magnetoresistance. The resistance can vary from more than 100 %. The Magnetic Random Access Memories (MRAM) and the read

heads of hard drive disks take benefits from this property. In order to increase storage capacity, a strong effort is made toward a size reduction of the MTJs. However, in the course of miniaturizing a unit cell for magnetic device a certain number of questions related to the finite size effect arise. Size effect starts then to play an important role and can affect drastically the electrodes magnetic properties. The results obtained on thin films cannot be straightforwardly extrapolated to patterned objects and nanostructures. Combining the high spatial resolution of XPEEM magnetic imaging, its chemical sensitivity and micromagnetic simulations, allowed to demonstrate the strong influence of a dipolar magnetic coupling on the magnetization reversal of MTJs [33]. Such studies are mandatory because understanding the domain wall formation process is crucial since it affects the magnetic – electrical properties of MTJs.

When the MTJs are structured in micrometer sized elements, an antiferromagnetic coupling tends to align the magnetization of each electrode in an antiparallel configuration contrary to what is usually observed in the thin film case. This coupling originates from the magnetic stray field at the nanostructures edges (see red arrows in Fig. 10). The MTJs is formed by two ferromagnetic layers: (permalloy and cobalt) separated by a thin aluminium oxide layer: Co(4 nm)/Al₂O₃(2 nm)/Fe₂₀Ni₈₀(4 nm). The MTJs have been patterned in ellipse-shaped structures in order to re-enforce the stray field effect. The high spatial resolution of the X-ray Photoemission Electron Microscopy (X-PEEM) combined to X-ray Magnetic Circular Dichroism (XMCD) enable to directly image the magnetic configuration of each electrode of the MTJs. This powerful technique allows to image independently the magnetic configuration in each electrodes if composed of different elements, thanks to its elemental selectivity. Figure 10a shows that, when no magnetic field is applied, both layer magnetizations are mainly in an antiparallel configuration. Moreover, if the Co layer magnetization is uniform, surprisingly this is not always the case in the NiFe layer. The (b) and (c) parts of Fig. 10 present images obtained in geometries where the technique is sensitive to the magnetization component either



Selected Synchrotron Radiation Techniques, Fig. 10 Top: Schematic representation of a MTJs. The two magnetic electrodes, composed of a nickel-iron alloy and of cobalt, are respectively represented in green and purple. They are separated by 2 nm thick layer of oxidized aluminium (not represented for clarity reason). The magnetic stray field sketched in red tends to align the magnetization each electrode in an antiparallel configuration. Bottom: XMCD-PEEM images of four ellipses recorded at the Co and at the Ni edges. The photons incidence direction is aligned along the ellipses long axis from the top left corner. The white and black contrasts correspond to magnetization components aligned along the long ellipses axis. XMCD-PEEM images of two ellipses (1 × 3 μm) measured at the Ni edge. The grey level distribution corresponds to the scalar projection of the local magnetization with respect to the light incidence: (b) parallel and (c) perpendicular to the ellipses long axis

along the ellipses long axis (b) or along the ellipses short axis (c). The combination of this imaging technique with *in silico* modelling (right part of the Fig. 2b, c) allows to understand the nature and the formation process of the non-uniform magnetization distribution presents in the NiFe layer. In this region, the magnetization rotates continuously by 360°, forming an object separating two regions of uniform magnetization. This magnetic object is called a 360° domain wall. By simulations, only one chirality is obtained for the wall, although three different cases can be observed experimentally: no wall and two chiralities (Fig. 10). This difference has been attributed to local magnetic anisotropy fluctuations at the ellipses extremities which drive the magnetization curling direction during the reversal process.

Future Directions for Research (Outlook)

Future challenges of GISAXS may lie in the use of the x-ray coherence [13, 18] in order to reconstruct single objects without any model assumptions. As far as the corresponding nano-objects can bear the X-ray beam, soft matter systems like polymers, block copolymers or even biological applications, such as proteins, peptides and viruses attached to surfaces or in lipid layers, have become a growing field unfolding the full potential of the GISAXS technique.

For μXRD experiments aiming at a characterization of individual micro and nano-structures [6, 23, 24], the actual demand is that of availability of brighter and smaller x-ray spots, with increased stability. Last generation synchrotron sources can fulfill in a certain measure these demands, and dedicated experimental stations are built [1]. Coupling μXRD with other techniques is achieved [6] and allows obtaining complementary results from various techniques, which completes the understanding about these systems and their properties. Another experimental path using focused X-ray beams is the one using a white (continuous energy spectrum) x-ray microbeam – the Laue microdiffraction experiments. It can give access to grain distributions (orientation, strain) in polycrystalline samples [34]. This particular topic needs

achromatic focusing elements (e.g., Kirkpatrick-Baez mirrors, capillaries) and was not detailed here. Alternative techniques like the lensless x-ray microscopy (Coherent Diffraction Imaging CDI [35], X-ray holography [36], ptychography [37], ...) using coherent x-ray beams are also a path opened for accessing non-destructively the properties of small structures: 3D imaging and strain determination with spatial resolutions of 10 nm and below were proved. Tremendous gain in photon flux and coherence can be achieved by the use of free electron laser sources. These approaches will not be detailed here; for the reader looking for more details, one can refer to the reviews [6, 38]. The non-invasive character of the XRD method and the relatively large penetration depth of the hard X-rays make possible experiments in which the diffraction approach is combined (in- or ex-situ) with other analysis methods. The particular cases of SEM and AFM have been examined. Ultra High Vacuum (UHV) environment with in-situ preparation facilities of thin metallic films was also addressed, but other environments (high temperature, clean environment, high pressure, reaction cells, electric or magnetic fields, ...) can easily be imagined. Experiments performed on model electronic devices while they were functioning [39] were also realized.

One of the main challenges for XPEEM microscope in the future is to combine the high spatial resolution with time-resolved experiments. In the case of magnetic materials for example, the dynamics of the magnetization reversal in thin magnetic films has become a matter of high interest for the future of magnetic recording and non-volatile magnetic memories. Parallel to the evolution towards smaller magnetic bits and memory cells, writing and reading times approaching the ns range will be required in a few years from now. A complete understanding of the magnetization dynamics in these structures requires the ability to probe the magnetization of the individual layers as well as their mutual interaction. Time-resolved X-PEEM measurements are very challenging, since the secondary electrons that are used for the image are strongly perturbed by the magnetic field necessary to switch the magnetization direction. Therefore,

the time-resolved XPEEM experiment can be only performed in stroboscopic mode. The magnetic pulses can be synchronized with the X-ray pulses coming from the storage ring to perform dynamic measurements in a pump-probe scheme. X-PEEM images cannot be acquired during the field pulses, but time resolution of 50 ps has been already demonstrated [40]. The advent of fourth generation synchrotron and more specifically X-ray Free Electron Laser [38] will certainly open up the way for high resolution imaging at a time scale of femtosecond.

Finally, many efforts are done to improve the spatial resolution. Although the ultimate resolution will be always limited by the diffraction limit (1 nm), the development of new aberration corrected setups will certainly open up the way to image nanostructure with resolution below 5 nm. Other aberrations correction methods using numerical algorithms or phase retrieval have been also proposed.

Cross-References

- AFM
- Nanoindentation
- SEM

References

1. See for example <http://www.lightsources.org/> for links to existing facilities
2. Wiedemann, H.: *Synchrotron Radiation*. Springer, Berlin (2002). ISBN 978-3540433927
3. Reimers, W., Pyzalla, A.R., Schreyer, A., Clements, H.: *Neutrons and Synchrotron Radiation in Engineering Materials Science: From Fundamentals to Material and Component Characterization*. Wiley-WCH/ GmbH & Co. KGaA, Weinheim (2008). ISBN 978-3-527-31533-8
4. Als-Nielsen, J., McMorrow, D.: *Elements of Modern X-Ray Physics*, 2nd edn. Wiley, New York (2010)
5. Barbier, A., Mocuta, C., Renaud, G.: Characterization and spectroscopy of thin films, Chapter 11. In: Nalwa, H.S. (ed.) *Handbook of Thin Film Materials*, vol. 2. Academic, San Diego (2002). ISBN 0-12-512910-6
6. Stangl, J., Mocuta, C., Chammard, V., Carbone, D.: *Nanobeam X-Ray Scattering: Probing Matter at the Nanoscale*. Wiley-VCH, Weinheim (2013). ISBN 978-3-527-41077-4
7. Locatelli, A., Bauer, E.: *Phys. Condens. Matter* **20**, 093002 (2008); Kuch, W.: *Imaging Magnetic Microspectroscopy*. Springer, Berlin (2003)
8. Bauer, E.: Low energy electron microscopy. *Rep. Prog. Phys.* **57**, 895–938 (1994). doi:10.1088/0034-4885/57/9/002
9. Riekel, C.: New avenues in x-ray microbeam experiments. *Rep. Prog. Phys.* **63**, 233 (2000); Davies, R.J., Burghammer, M., Riekel, C.: A combined microRaman and microdiffraction set-up at the European Synchrotron Radiation Facility ID13 beamline. *J. Synchr. Rad.* **16**, 22 (2009)
10. Renaud, G., Lazzari, R., Leroy, F.: Probing surface and interface morphology with grazing incidence small angle X-Ray scattering. *Sur. Sci. Rep.* **64**, 255–380 (2009)
11. Rauscher, M., et al.: Grazing incidence small angle x-ray scattering from free-standing nanostructures". *J. Appl. Phys.* **86**, 6763 (1999). doi:10.1063/1.371724
12. Barbier, A., Stanescu, S., Boeglin, C., Deville, J.-P.: Local morphology and correlation lengths of reactive NiO/Cu(111) interfaces. *Phys. Rev. B* **68**, 245418-1-7 (2003)
13. Zozulya, A.V., Yefanov, O.M., Vartanyants, I.A., Mundboth, K., Mocuta, C., Metzger, T.H., Stangl, J., Bauer, G., Boeck, T., Schmidbauer, M.: Imaging of nanoislands in coherent grazing-incidence small-angle x-ray scattering experiments. *Phys. Rev. B* **78**, 121304 (2008)
14. Mane Mane, J., Cojocaru, C.S., Barbier, A., Deville, J. P., Jean, B., Metzger, T.H., Thiodjo Sendja, B., Le Normand, F.: GISAXS study of carbon nanotubes grown on SiO₂/Si(100) by CVD. *Phys. Status Solidi (RRL)* **1**, 122–124 (2007)
15. Gilles, R., Rémi, L., Christine, R., Antoine, B., Marion, N., Olivier, U., Frédéric, L., Jacques, J., Yves, B., Henry, C.R., Jean-Paul, D., Fabrice, S., Jeannot, M.-M., Olivier, F.: Real-time monitoring of growing nanoparticles. *Science* **300**, 1416 (2003)
16. Papadakis, C.M., Di, Z., Posselt, D., Smilgies, D.-M.: Structural instabilities in Lamellar Diblock Copolymer thin films during solvent vapor uptake. *Langmuir* **24**, 13815–13818 (2008)
17. Smilgies, D.-M., Li, R., Giri, G., Chou, K.W., Diao, Y., Bao, Z., Amassian, A.: Look fast: crystallization of conjugated molecules during solution shearing probed in-situ and in real time by X-ray scattering. *Phys. Status Solidi (RRL)* **7**, 177–179 (2013)
18. Sun, T., Jiang, Z., Strzalka, J., Ocola, L., Wang, J.: Three-dimensional coherent X-ray surface scattering imaging near total external reflection. *Nat. Photonics* **6**, 586–590 (2012)
19. Zhang, J., Posselt, D., Sepe, A., Shen, X., Perlich, J., Smilgies, D.-M., Papadakis, C.M.: Structural evolution of perpendicular Lamellae in Diblock Copolymer thin films during solvent vapor treatment investigated by grazing-incidence small-angle X-Ray scattering. *Macromol. Rapid Commun.* **34**, 1289–1295 (2013)

20. Senesi, A.J., Eichelsdoerfer, D.J., Macfarlane, R.J., Jones, M.R., Auyeung, E., Lee, B., Mirkin, C.A.: Stepwise evolution of DNA-programmable nanoparticle superlattices. *Angew. Chem. Int. Ed.* **52**, 6624–6628 (2013)
21. Perlich, J., Schwartzkopf, M., Körstgens, V., Erb, D., Risch, J.F.H., Müller-Buschbaum, P., Röhlsberger, R., Roth, S.V., Gehrke, R.: Pattern formation of colloidal suspensions by dip-coating: an *in situ* grazing incidence X-ray scattering study. *Phys. Status Solidi (RRL)* **6**, 253–255 (2012)
22. Schmidbauer, M.: X-Ray Diffuse Scattering from Self Organized Mesoscopic Semiconductor Structures. Springer Tracts in Modern Physics, vol. 199. Springer, Berlin (2004)
23. Dubslaff, M., Hanke, M., Schoder, S., et al.: X-ray nanodiffraction at individual SiGe/Si(001) dot molecules and its numerical description based on kinematical scattering theory. *Appl. Phys. Lett.* **96**(13), 133107 (2010)
24. Mocuta, C., Stangl, J., Mundboth, K., et al.: Beyond the ensemble average: x-ray microdiffraction analysis of single SiGe island. *Phys. Rev. B* **77**, 245425 (2008)
25. Mocuta, C., Barbier, A., Ramos, A.V., et al.: Effect of optical lithography patterning on the crystalline structure of tunnel junctions. *Appl. Phys. Lett.* **91**(24), 241917 (2007)
26. Mocuta, C., Barbier, A., Stanescu, S., et al.: X-ray diffraction imaging of metal-oxide epitaxial tunnel junctions made by optical lithography: use of focused and unfocused X-ray beams. *J. Synchrotron Radiat.* **20**, 355–365 (2013)
27. Ren, Z., Mastropietro, F., Davydok, A., et al.: Scanning force microscope for *in situ* nanofocused X-ray diffraction studies. *J. Synchrotron Radiat.* **21**, 1128–1133 (2014)
28. Cornelius, T.W., Mastropietro, F., Thomas, O., Schulli, T.U.: In *situ* nanofocused X-ray diffraction combined with scanning probe microscopy, Chapter 10. In: Shih, K. (ed) X-Ray Diffraction: Structure, Principle and Applications, New York, USA, pp. 233–259. Nova Science (2013)
29. Fink, R., et al.: *J. Electron Spectrosc. Relat. Phenom.* **84**, 1249 (1997); Feng, J., et al.: *J. Phys. Condens. Matter* **17**, S1339 (2005)
30. Stöhr, J., et al.: Element-specific Magnetic Microscopy with Circularly Polarized X-rays. *Science* **259**, 658 (1993)
31. Ohldag, H., et al.: Spin Reorientation at the Antiferromagnetic NiO(001) Surface in Response to an Adjacent Ferromagnet. *Phys. Rev. Lett.* **86**, 2878 (2001)
32. Bauer, E., Belkhou, R., Cherifi, S., Locatelli, A., Pavlovska, A., Rougemaille, N.: Magnetostructure of MnAs on GaAs revisited. *J. Vac. Sci. Technol. B* **25**, 1470 (2007)
33. Hehn, M., et al.: 360° domain wall generation in the soft layer of magnetic tunnel junctions. *Appl. Phys. Lett.* **92**, 072501 (2008)
34. Budai, J.D., Yang, W., Tamura, N., et al.: X-ray microdiffraction study of growth modes and crystallographic tilts in oxide films on metal substrates. *Nat. Mater.* **2**, 487 (2003)
35. Beutier, G., Verdier, M., Parry, G., et al.: Strain inhomogeneity in copper islands probed by coherent X-ray diffraction. *Thin Solid Films* **530**, 120–124 (2013)
36. Chamard, V., Stangl, J., Carbone, D., et al.: Three-dimensional X-ray Fourier transform holography: the Bragg case. *Phys. Rev. Lett.* **104**, 165501 (2010)
37. Godard, P., Carbone, D., Alain, M., et al.: Three dimensional X-ray Bragg ptychography: high resolution imaging of extended crystalline nanostructures. *Nat. Commun.* **2**, 568 (2011)
38. Vartanyants, I.A., Robinson, I.K., McNulty, I., et al.: Coherent scattering and lensless imaging at the European XFEL Facility. *J. Synchrotron Radiat.* **14**, 453 (2007)
39. Hrauda, N., Zhang, J., Wintersberger, E., et al.: X-ray nanodiffraction on a single SiGe quantum dot inside a functional field-effect transistor. *Nano Lett.* **11**(7), 2875–2880 (2011)
40. Choe, S.-B., et al.: Vortex Core–Driven Magnetization Dynamics. *Science* **304**, 420 (2004)

Self-Assembled Monolayers

► Nanostructures for Surface Functionalization and Surface Properties

Self-Assembled Monolayers for Nanotribology

Bharat Bhushan
Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Synonyms

Molecularly thick layers; Monolayer lubrication; Organic films; Self-organized layers

Definition

Organized and dense molecular-scale layers of long-chain, organic molecules are referred to as

self-assembled monolayers (SAMs). SAMs are molecularly thick, well-organized, and chemically bonded to the substrate. Ordered molecular assemblies with various properties can be engineered using chemical grafting of various polymer molecules with suitable functional head groups, spacer chains, and surface terminal groups.

Overview

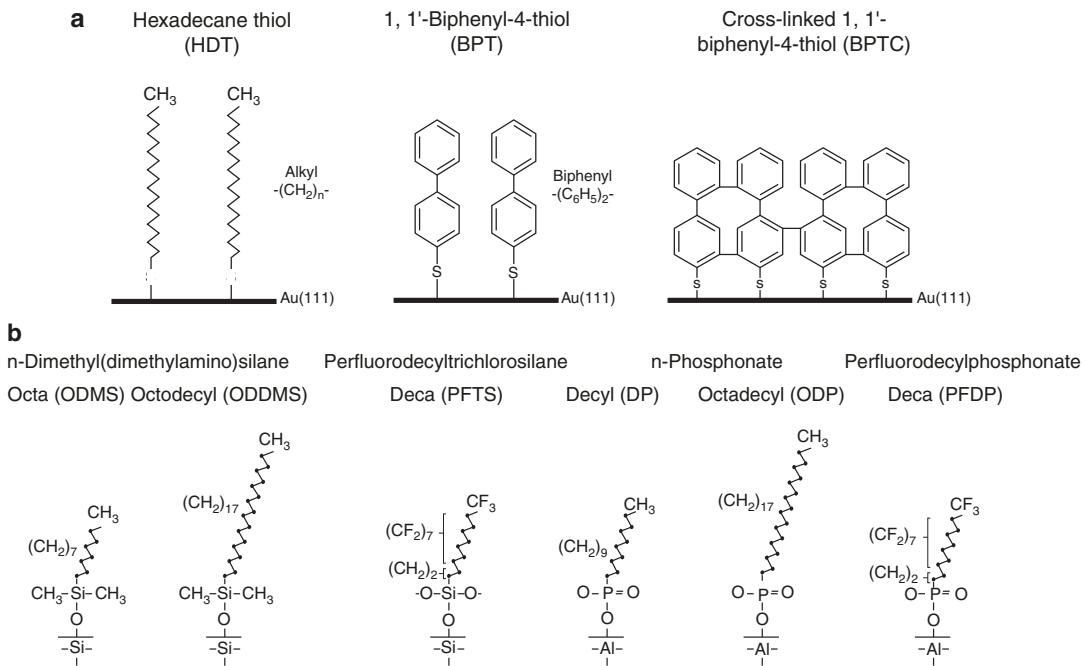
Reliability of various micro- and nanodevices, also commonly referred to as micro/nanoelectromechanical systems (MEMS/NEMS) and BioMEMS/BioNEMS, requiring relative motion, as well as magnetic storage devices (which include magnetic rigid disk and tape drives) requires the use of hydrophobic and lubricating films to minimize adhesion, stiction, friction, and wear [2–4, 7, 9–12, 15, 36, 39]. In various applications, surfaces need to be protected from exposure to the operating environment. For example, in various biomedical applications, such as biosensors and implantable biomedical devices, undesirable protein adsorption, biofouling, and biocompatibility are some of the major issues [9–11]. In micro- and nanofluidic based sensors, the fluid drag in micro- and nanochannels can be reduced by using hydrophobic coatings [42]. Selected hydrophobic films are needed for these applications.

For lubrication, an effective approach involves the deposition of organized and dense molecular layers of long-chain molecules. Two common methods to produce monolayers and thin films are the Langmuir-Blodgett (L-B) deposition and self-assembled monolayers (SAMs) by chemical grafting of molecules. LB films are physically bonded to the substrate by weak van der Waals attraction, while SAMs are chemically bonded via covalent bonds to the substrate. Because of the choice of chain length and terminal linking group that SAMs offer, they hold great promise for boundary lubrication of MEMS/NEMS. A number of studies have been conducted to study tribological properties of various SAMs deposited on Si, Al, and Cu substrates [14, 16–18, 21, 24–31, 33–35, 38, 40].

Bhushan and Liu [14], Liu et al. [35], and Liu and Bhushan [8] studied adhesion, friction and wear properties of alkylthiol and biphenylthiol SAMs on Au(111) films. They explained the friction mechanisms using a molecular spring model in which local stiffness and intermolecular forces govern the friction properties. They studied the influence of relative humidity, temperature, and velocity on adhesion and friction. They also investigated the wear mechanisms of SAMs by a continuous microscratch AFM technique.

Fluorinated carbon (fluorocarbon) molecules are known to have low surface energy and are commonly used for lubrication [5, 7, 8, 12]. Bhushan and Cichomski [13] deposited fluorosilane SAMs on polydimethylsiloxane (PDMS). To make a hydrophobic PDMS surface chemically active, PDMS surface was oxygenated using an oxygen plasma, which introduces silanol groups (SiOH). They reported that SAM coated PDMS was more hydrophobic with lower adhesion, friction, and wear. Bhushan et al. [17, 19], Kasai et al. [30], Lee et al. [31], Tambe and Bhushan [38], and Tao and Bhushan [41] studied the adhesion, friction, and wear of methyl- and/or perfluoro- terminated alkylsilanes on silicon. They reported that perfluoroalkylsilane SAMs exhibited lower surface energy, higher contact angle, lower adhesive force, and lower wear as compared to that of alkylsilanes. Kasai et al. [30] also reported the influence of relative humidity, temperature, and velocity on adhesion and friction. Tao and Bhushan [40] studied degradation mechanisms of alkylsilanes and perfluoroalkylsilane SAMs on Si. They reported that oxygen in the air causes thermal oxidation of SAMs.

Tambe and Bhushan [38], Bhushan et al. [18], Hoque et al. [24, 25] and DeRose et al. [21] studied the nanotribological properties of methyl- and perfluoro-terminated alkylphosphonate, perfluorodecyldimethylchlorosilane, and perfluorodecanoic acid on aluminum, of industrial interest. Hoque et al. [26] and DeRose et al. [21] studied the nanotribological properties of alkylsilanes and perfluoroalkylsilanes on aluminum. Hoque et al. [27–29] studied the nanotribological properties of alkylphosphonate and perfluoroalkylsilane SAMs on copper. The



Self-Assembled Monolayers for Nanotribology,

Fig. 1 Schematics of the structures of (a) hexadecane and biphenyl thiol SAMs on Au(111) substrates, and (b)

perfluoroalkylsilane and alkylsilane SAMs on Si with native oxide substrates, and perfluoroalkylphosphonate and alkylphosphonate SAMs on Al with native oxide

authors found that these SAMs on aluminum and copper perform well irrespective of the substrate used. They confirmed the presence of respective films using X-ray photoelectron spectroscopy (XPS).

Hoque et al. [27–29] studied the chemical stability of various SAMs deposited on Cu substrates via exposure to various corrosive conditions. DeRose et al. [21] studied the chemical stability of various SAMs deposited on Al substrates via exposure to corrosive conditions (aqueous nitric acid solutions of a low pH of 1.8 at temperatures ranging from 60 °C to 80 °C for times ranging from 30 to 70 min). The exposed samples were characterized by XPS and contact angle measurements. They reported that perfluorodecanoic acid/Al is less stable than perfluorodecylphosphonate/Al and octadecylphosphonate/Al, but more stable than perfluorodecyldimethylchlorosilane/Al, which has implications in digital micromirror devices (DMD) applications. In general, chemical stability data of various SAMs deposited on Cu and Al surfaces to corrosive environments has

been reported by these authors. Based on these studies, it was concluded that chemisorption occurs at the interface and is responsible for strong interfacial bonds.

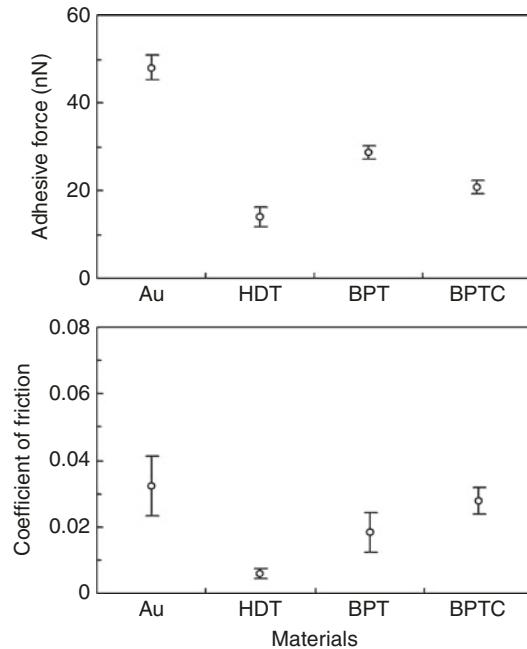
To date, the contact angle and nanotribological properties of alkanethiol, biphenylthiol, alkylsilane, perfluoroalkylsilane, alkylphosphonate and perfluoroalkylphosphane SAMs have been widely studied. In the following, the nanotribological properties of various SAMs are reviewed having alkyl and biphenyl spacer chains with different surface terminal groups ($-\text{CH}_3$, $-\text{CF}_3$) and head groups ($-\text{S}-\text{H}$, $-\text{Si}-\text{O}-$, $-\text{OH}$, and $\text{P}-\text{O}-$) which have been investigated by AFM at various operating conditions [14, 17, 18, 30, 33–35, 38, 40]. Hexadecane thiol (HDT), 1, 1'-biphenyl-4-thiol (BPT), and crosslinked BPT (BPTC) were deposited on Au(111) films on Si(111) substrates by immersing the substrate in a solution containing the precursor (ligand) that is reactive to the substrate surface (Fig. 1a). Crosslinked BPTC was produced by irradiation of BPT monolayers with low energy electrons. Perfluoroalkylsilane and alkylsilane

SAMs were deposited on Si(100) by exposing the substrate to the vapor of the reactive chemical precursors (Fig. 1b). Perfluoroalkylphosphonate and alkylphosphonate SAMs were deposited on sputtered Al film on Si substrate as well as bulk Al substrates (Fig. 1b).

Hexadecane Thiol and Biphenyl Thiol SAMs on Au(111)

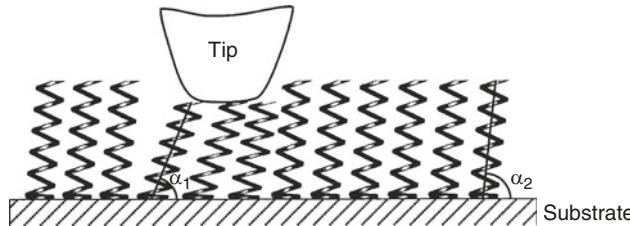
Bhushan and Liu [14] studied the effect of film compliance on adhesion and friction. They used hexadecane thiol (HDT), 1,1,biphenyl-4-thiol (BPT), and crosslinked BPT (BPTC) solvent deposited on Au(111) substrate, Fig. 1a. The average values and standard deviation of the adhesive force and coefficient of friction are presented in Fig. 2. Based on the data, the adhesive force and coefficient of friction of SAMs are less than the corresponding substrates. Among various films, HDT exhibits the lowest values. Based on stiffness measurements of various SAMs, HDT was most compliant, followed by BPT and BPTC. Based on friction and stiffness measurements, SAMs with high-compliance long carbon chains exhibit low friction; chain compliance is desirable for low friction. Friction mechanism of SAMs is explained by a so-called “molecular spring” model (Fig. 3). According to this model, the chemically adsorbed self-assembled molecules on a substrate are just like assembled molecular springs anchored to the substrate. An asperity sliding on the surface of SAMs is like a tip sliding on the top of “molecular springs or brush.” The molecular spring assembly has compliant features and can experience orientation and compression under load. The orientation of the molecular springs or brush under normal load reduces the shearing force at the interface, which in turn reduces the friction force. The orientation is determined by the spring constant of a single molecule as well as the interaction between the neighboring molecules, which can be reflected by packing density or packing energy. It should be noted that the orientation can lead to conformational defects along the molecular chains, which lead to energy dissipation.

An elegant way to demonstrate the influence of molecular stiffness on friction is to investigate



Self-Assembled Monolayers for Nanotribology,
Fig. 2 Adhesive forces and coefficients of friction of Au(111) and various SAMs

SAMs with different structures on the same wafer. For this purpose, a micropatterned SAM was prepared. First the biphenyldimethylchlorosilane (BDCS) was deposited on silicon by a typical self-assembly method [33]. Then the film was partially crosslinked using mask technique by low energy electron irradiation. Finally the micropatterned BDCS films were realized, which had the as-deposited and crosslinked coating regions on the same wafer. The local stiffness properties of this micropatterned sample were investigated by force modulation AFM technique [22]. The variation in the deflection amplitude provides a measure of the relative local stiffness of the surface. Surface height, stiffness, and friction images of the micropatterned biphenyldimethylchlorosilane (BDCS) specimen are obtained and presented in Fig. 4 [33]. The circular areas correspond to the as-deposited film, and the remaining area to the crosslinked film. Figure 4a indicates that crosslinking caused by the low energy electron irradiation leads to about 0.5 nm decrease of the surface height of BDCS films. The corresponding stiffness images indicate that the



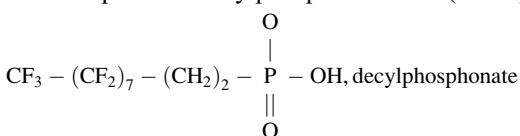
Self-Assembled Monolayers for Nanotribology,
Fig. 3 Molecular spring model of SAMs. In this figure, $\alpha_1 < \alpha_2$, which is caused by the orientation under the normal load applied by AFM tip. The orientation of the molecular springs reduces the shearing force at the

crosslinked area has higher stiffness than the as-deposited area. Figure 4b indicates that the as-deposited area (higher surface height area) has a lower friction force. Obviously, these data of the micropatterned sample prove that the local stiffness of SAMs has an influence to their friction performance. Higher stiffness leads to larger friction force. These results provide a strong proof of the suggested molecular spring model.

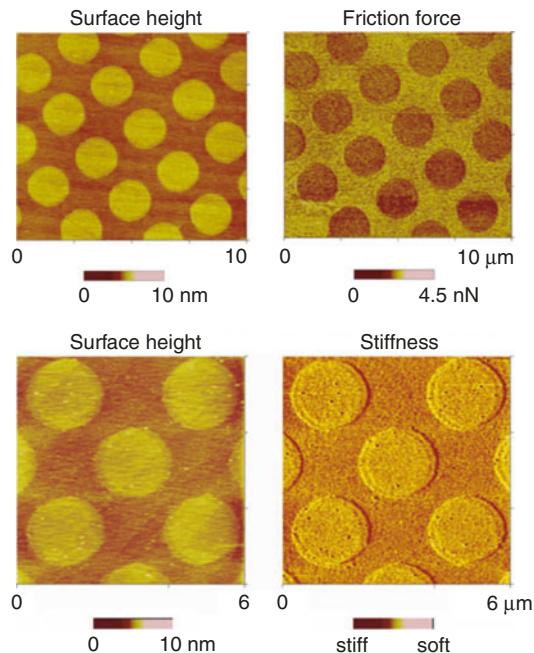
The SAMs with high-compliance long carbon chains also exhibit the best wear resistance [14, 33]. In wear experiments, the wear depth as a function of normal load curves show a critical normal load, at which film wears rapidly. A representative curve is shown in Fig. 5. Below the critical normal load, SAMs undergo orientation; at the critical load SAMs wear away from the substrate due to relatively weak interface bond strengths, while above the critical normal load severe wear takes place on the substrate.

Perfluoroalkylsilane and Alkylsilane SAMs on Si(100), and Perfluoroalkylphosphonate and Alkylphosphonate SAMs on Al

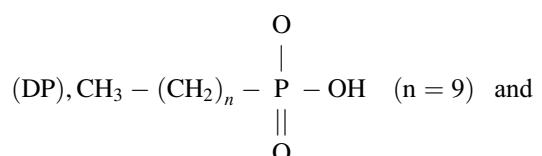
Perfluorodecyltricholorosilane (PFTS), $\text{CF}_3-(\text{CF}_2)_7-(\text{CH}_2)_2-\text{SiCl}_3$, n-octyldimethyl (dimethyl amino)silane (ODMS), $\text{CH}_3-(\text{CH}_2)_n-\text{Si}(\text{CH}_3)_2-\text{N}(\text{CH}_3)_2$ ($n = 7$), and n-octadecyldimethyl(dimethylamino)silane ($n = 17$) (ODDMS) vapor deposited on Si(100) substrate and perfluorodecylphosphonate (PFDP)



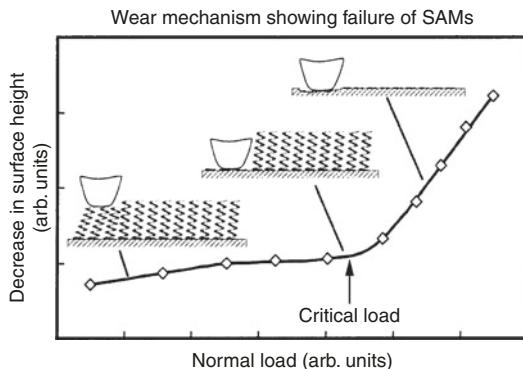
interface, which in turn reduces the friction force. The molecular spring constant, as well as the inter-molecular forces can determine the magnitude of the coefficients of friction of SAMs. In this figure, the size of the tip and molecular springs do not in the exactly scale [14]



Self-Assembled Monolayers for Nanotribology,
Fig. 4 (a) AFM Grayscale surface height and stiffness images, and (b) AFM grayscale surface height and friction force images of micropatterned BDCS [33]



octadecylphosphonate (ODP) ($n = 17$) by liquid deposition on sputtered Al film on Si substrate were selected (Fig. 1b). Perfluoro SAMs were selected because fluorinated films are known to have low surface energy. Two chain lengths of



Self-Assembled Monolayers for Nanotribology,
Fig. 5 Illustration of the wear mechanisms of SAMs with increasing normal load [33]

alkylsilanes (with 8 and 18 carbon atoms) were selected to compare their nanotribological performance with that of the former as well as to study the effect of chain length. Al substrate was selected because of the application of Al micromirrors in digital projection displays. Perfluoroalkylphosphane (with ten carbon atoms) and alkylphosphonate SAMs (with 10 and 18 carbon atoms) on Al were selected.

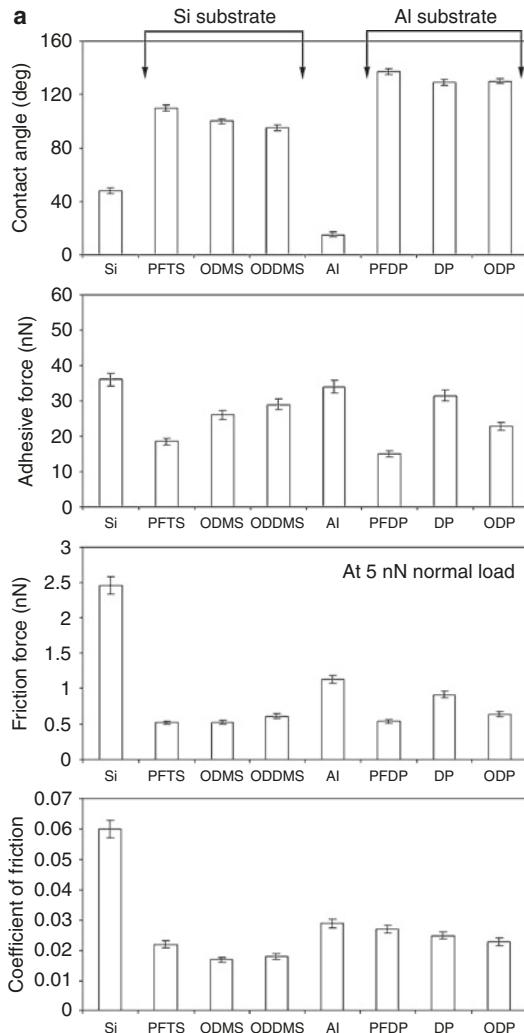
The measured values are compared among the samples in Fig. 6a [17, 18]. Significant improvement in the water repellent property was observed for perfluorinated SAMs as compared to bare Si and Al substrates. Static contact angles of alkylsilanes and alkylphosphonates were also higher than corresponding substrates, but lower than corresponding perfluorinated films. The contact angle generally increases with a decrease in surface energy [23], which is consistent with the data obtained. The contact angles can be influenced by the packing density as well as the sample roughness [37]. The higher contact angles for the SAMs deposited on Al substrates than those on Si substrate are probably due to this effect. The $-CH_3$ groups in ODMS, ODDMS, DP, and ODP are non-polar and are known to contribute to the water repellent property. Perfluorinated SAMs exhibited the highest contact angle among the SAMs tested in this study.

Figure 6a shows the adhesive force, friction force, and the coefficient of friction measured under ambient conditions using an AFM, and

Fig. 6b shows the friction force vs. normal load plots for various SAMs deposited onto the Si and Al substrates [17, 18]. The bare substrates showed higher adhesive force than the SAMs coatings. ODMS and ODDMS shows an adhesive force comparable to DP and ODP despite their lower water contact angles. These SAMs have the same tail groups, and during AFM measurements the AFM tip interacts only with the tail groups, whereas the contact angles can also be influenced by the head groups in these SAMs. This is probably the reason as to why the adhesive forces for these SAMs are comparable. PFTS and PFDP, which have the highest contact angles, showed the lowest adhesion.

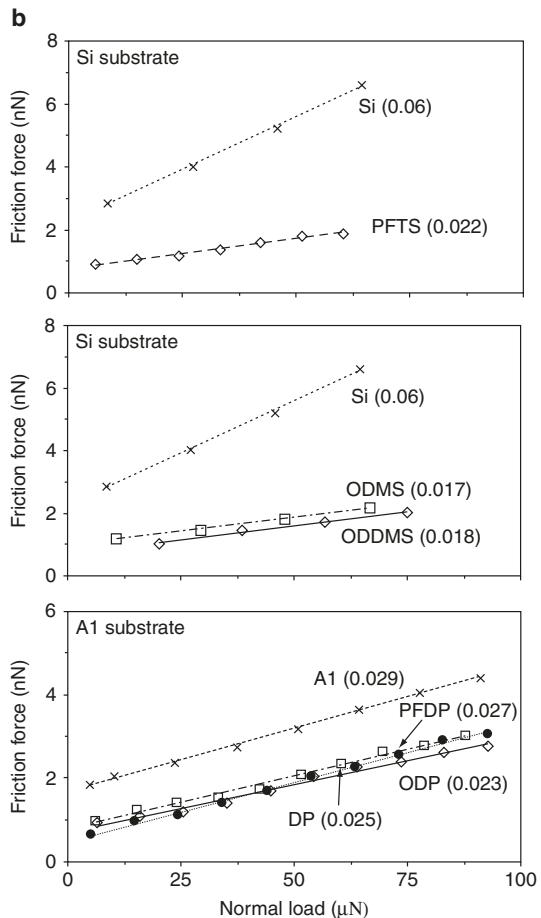
The effect of relative humidity for various SAMs on adhesion and friction was studied. Adhesive force, friction force at 5 nN of normal load, coefficient of friction, and microwear data are presented in Fig. 7 [18, 30]. The result of adhesive force for silicon showed an increase with relative humidity, Fig. 7. This is expected since the surface of silicon is hydrophilic, as shown in Fig. 6a. More condensation of water at the tip-sample interface at higher humidity increases the adhesive force due to capillary effect. On the other hand, the adhesive force for the SAMs showed a very weak dependency on the change in humidity. This occurs since the surface of the SAMs is hydrophobic. The adhesive force of ODMS/Si and ODDMS/Si showed a slight increase from 75 % to 90 % RH. Such an increase was absent for PFTS/Si, possibly because of the hydrophobicity of PFTS/Si.

The friction force of silicon showed an increase with relative humidity up to about 75 % RH and a slight decrease beyond this point, see Fig. 7. The initial increase can result from the increase in adhesive force. The decrease in friction force at higher humidity could be attributed to the lubricating effect of the water layer. This effect is more pronounced in the coefficient of friction. Since the adhesive force increased and coefficient of friction decreased in this range, those effects cancel each other out and the resulting friction force showed slight changes. On the other hand, the friction force and coefficient of friction of SAMs showed very small changes with relative humidity



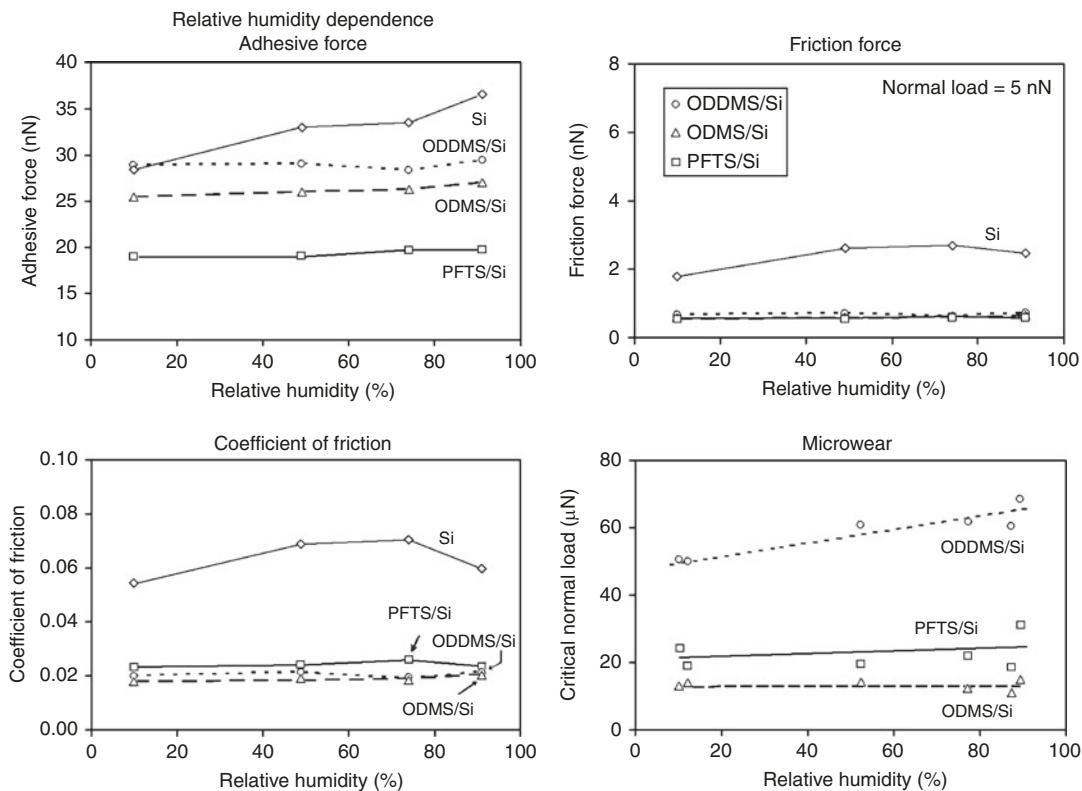
Self-Assembled Monolayers for Nanotribology, Fig. 6 (a) The static contact angle, adhesive force, friction force and coefficient of friction measured using an AFM

like that found for adhesive force. This suggests that the adsorbed water layer on the surface maintained a similar thickness throughout the relative humidity range tested. The differences among the SAM types were small within the measurement error, however a closer look at the coefficient of friction for ODMS/Si showed a slight increase from 75 % to 90 % RH as compared to PFTS/Si, possibly due to the same reason for the adhesive force increment. The inherent hydrophobicity of SAMs means that they did not show much relative humidity dependence.



for various SAMs on Si and Al substrates, and (b) friction force vs. normal load plots for various SAMs on Si and Al substrates [17, 18]

Figure 8 shows the effect of temperature on adhesive force, friction force at 5 nN of normal load, and coefficient of friction for various SAMs on Si substrate [30]. The adhesive force showed an increase with the temperature, from room temperature (RT) to about 55 °C, followed by a decrease from 55 °C to 75 °C, and eventually leveled off from 75 °C to 100 °C. The initial increase of adhesive force at lower temperatures is not well understood. The observed decrease could be attributed to the desorption of water molecules on the surface. After almost full depletion of the water



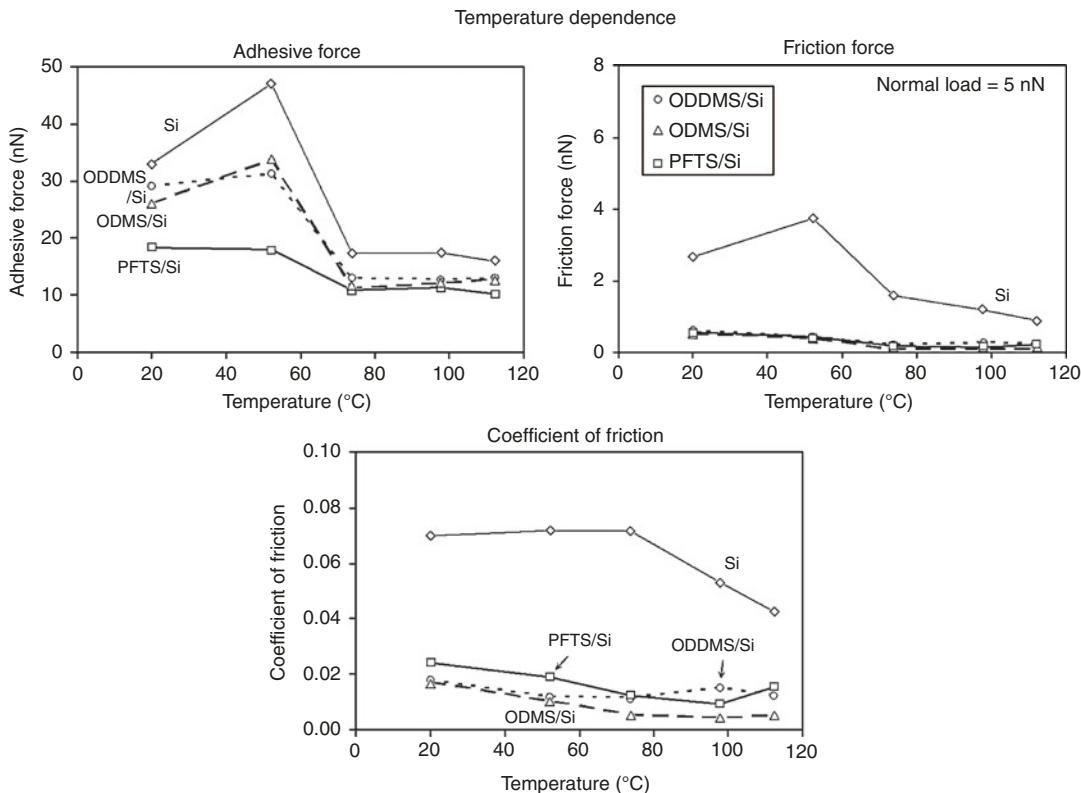
Self-Assembled Monolayers for Nanotribology, Fig. 7 Relative humidity effect on adhesive force, friction force, coefficient of friction and microwear for various SAMs on Si substrates [30]

layer, the adhesive force remains constant. The SAMs with hydrocarbon backbones showed similar behavior as that of the Si substrates, but the initial increase in the adhesive force with temperature was smaller. The SAMs with fluorocarbon backbone chains showed almost no temperature dependence. For the SAMs with hydrocarbon backbone chains, the initial increase in adhesive force is believed to be caused by the melting of the SAM film. The melting point for a linear carbon chain molecule such as $\text{CH}_3(\text{CH}_2)_{14}\text{CH}_2\text{OH}$ is 50 °C [32]. With an increase in temperature, the SAM film softens, thereby increasing the real area of contact and consequently the adhesive force. Once the temperature is higher than the melting point, the lubrication regime is changed from boundary lubrication in a solid SAM to liquid lubrication in the melted SAM [33].

The friction force for silicon showed an increase with temperature followed by a steady

decrease. The friction force is highly affected by the change in adhesion. The decrease in friction can result from the depletion of the water layer. The coefficient of friction for silicon remained constant followed by a decrease starting at about 80 °C. For SAMs, the coefficient of friction exhibited a monotonic decrease with temperature. The decrease in friction and coefficient of friction for SAMs possibly results from the decrease in stiffness. As introduced before, the spring model suggests a smaller friction for more compliant SAMs [33]. The difference among the SAM types was not significant. PFTS could maintain its stiffness more than ODMS and ODDMS when temperature is increased [20], however, it was not pronounced in the results.

Figure 9a shows the relationship between the decrease in surface height as a function of the normal load during wear tests [18, 30]. As shown in the figure, the SAMs exhibit a critical



Self-Assembled Monolayers for Nanotribology, Fig. 8 Temperature effect on adhesive force, friction force, and coefficient of friction for various SAMs on Si substrates [30]

normal load, beyond the point of which the surface height drastically decreases. Figure 9a also shows the wear behavior of the Al and Si substrates. Unlike the SAMs, the substrates show a monotonic decrease in surface height with the increasing normal load with wear initiating from the very beginning, i.e., even for low normal loads. Si (Young's modulus of elasticity, $E = 130 \text{ GPa}$ [1], hardness, $H = 11 \text{ GPa}$ [6]) is relatively hard in comparison to Al ($E = 77 \text{ GPa}$, $H = 0.41 \text{ GPa}$), and hence the decrease in surface height for Al is much larger than that for Si for similar normal loads.

The critical loads corresponding to the sudden failure of SAMs are shown in Fig. 9b. Amongst all the SAMs, ODDMS shows the best performance in the wear tests, and this is believed to be because of the longer chain length effect. Fluorinated SAMs – PFTS and PFDP show a higher critical load as compared to ODMS and DP with similar

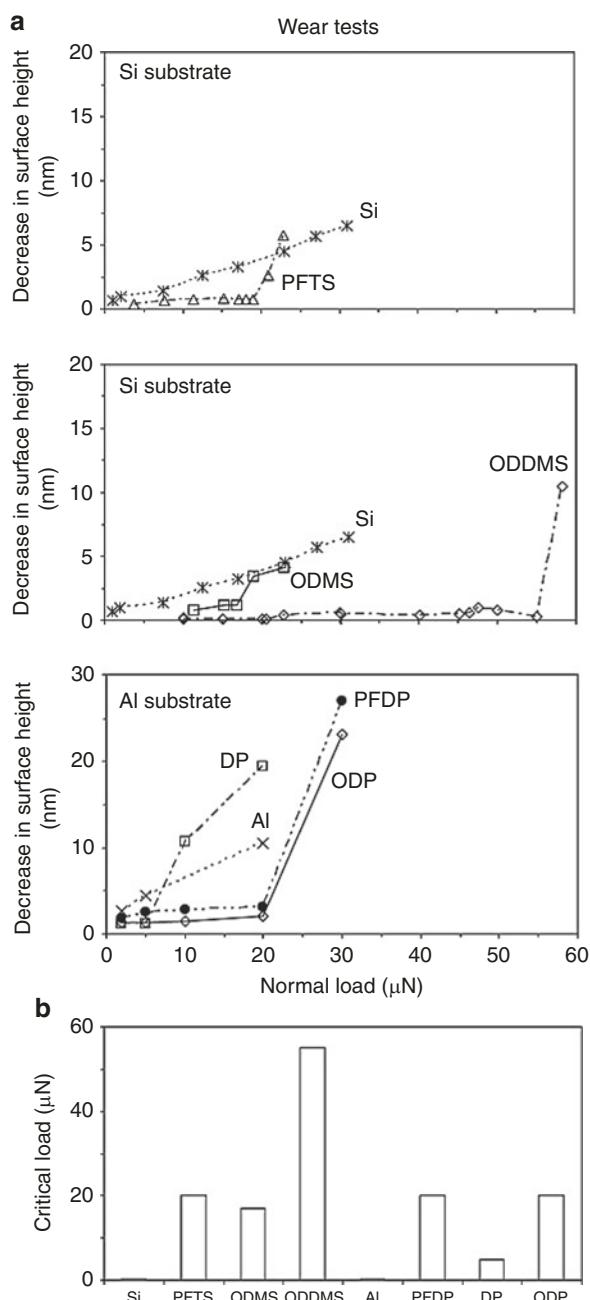
chain lengths. ODP shows a higher critical load as compared to DP because of its longer chain length. The mechanism of failure of compliant SAMs during wear tests has been presented earlier in Fig. 5. It is believed that the SAMs fail mostly due to shearing of the molecule at the head group, that is, by means of shearing of the molecules off the substrate.

Closure

The contact angle, adhesion, friction, and wear properties of SAMs having alkyl, biphenyl, and perfluoroalkyl spacer chains with different surface terminal groups ($-\text{CH}_3$ and $-\text{CF}_3$) and head groups ($-\text{SH}$, $-\text{Si}-\text{O}-$, $-\text{OH}$, and $\text{P}-\text{O}-$) studied are presented in this article. It is found that the adhesive force varies linearly with the work of adhesion value of SAMs, which indicates that capillary condensation of water plays an important role to the adhesion of SAMs on the

Self-Assembled Monolayers for Nanotribology

Fig. 9 (a) Decrease in surface height as a function of normal load after one scan cycle for various SAMs on Si and Al substrates, and (b) comparison of critical loads for failure during wear tests for various SAMs [18, 30]



nanoscale at ambient conditions. SAMs with high-compliance long carbon spacer chains exhibit the lowest adhesive force and friction force. The friction data are explained using a molecular spring model, in which the local stiffness and intermolecular force govern frictional performance. The results of the stiffness and

friction characterization of the micropatterned sample with different structures support this model. Perfluoroalkylsilane and perfluoroalkylphosphonate SAMs exhibit lower surface energy, higher contact angle, and lower adhesive force as compared to that of alkylsilane and alkylphosphonate SAMs, respectively. The substrate had

little effect. The coefficient of friction of various SAMs were comparable.

The influence of relative humidity on adhesion and friction of SAMs is dominated by the thickness of the adsorbed water layer. At higher humidity, water increases friction through increased adhesion by meniscus effect in the contact zone. With an increase in temperature, the desorption of the adsorbed water layer and reduction of the surface tension of water reduces the adhesive force and friction force. A decrease in adhesion and friction with temperature was found for all films.

PFTS/Si showed a better wear resistance than ODMS/Si. ODDMS/Si showed a better wear resistance than ODMS/Si due to the chain length effect. Wear behavior of the SAMs is mostly determined by the molecule-substrate bond strengths. Similar trends were observed for films on Al substrates.

Cross-References

- [Nanotechnology](#)
- [Nanotribology](#)
- [Reliability of Nanostructures](#)

References

1. Anonymous: Properties of silicon, EMIS data reviews series No. 4, INSPEC, Institution of Electrical Engineers, London (1988) (see also Anonymous, MEMS materials database, <http://www.memsnet.org/material/> (2002))
2. Bhushan, B.: Tribology and Mechanics of Magnetic Storage Devices, 2nd edn. Springer, New York (1996)
3. Bhushan, B. (ed.): Tribology Issues and Opportunities in MEMS. Kluwer, Dordrecht (1998)
4. Bhushan, B. (ed.): Handbook of Micro/Nanotribology, 2nd edn. Boca Raton, CRC Press (1999)
5. Bhushan, B.: Principles and Applications of Tribology. Wiley, New York (1999)
6. Bhushan, B.: Chemical, mechanical and tribological characterization of ultra-thin and hard amorphous carbon coatings as thin as 3.5 nm: recent developments. *Diam. Relat. Mater.* **8**, 1985–2015 (1999)
7. Bhushan, B. (ed.): Modern Tribology Handbook. Principles of Tribology, Vol. 1. Materials, Coatings, and Industrial Applications, vol. 2. CRC Press, Boca Raton (2001)
8. Bhushan, B.: Introduction to Tribology. Wiley, New York (2002)
9. Bhushan, B.: Nanotribology and nanomechanics of MEMS/NEMS and BioMEMS/BioNEMS materials and devices. *Microelectron. Eng.* **84**, 387–412 (2007)
10. Bhushan, B.: Nanotribology and nanomechanics in nano/biotechnology. *Philos. Tr. R. Soc. A* **366**, 1499–1537 (2008)
11. Bhushan, B.: Springer Handbook of Nanotechnology, 3rd edn. Springer, Heidelberg (2010)
12. Bhushan, B.: Nanotribology and Nanomechanics I – Measurement Techniques and Nanomechanics, II – Nanotribology, Biomimetics, and Industrial Applications, 3rd edn. Springer, Heidelberg (2011)
13. Bhushan, B., Cichomski, M.: Nanotribological characterization of vapor phase deposited fluorosilane self-assembled monolayers deposited on polydimethylsiloxane surfaces for biomedical micro-/nanodevices. *J. Vac. Sci. Technol. A* **25**, 1285–1293 (2007)
14. Bhushan, B., Liu, H.: Nanotribological properties and mechanisms of alkylthiol and biphenyl thiol self-assembled monolayers studied by atomic force microscopy. *Phys. Rev. B* **63**, 245412-1–245412-11 (2001)
15. Bhushan, B., Israelachvili, J.N., Landman, U.: Nanotribology: friction, wear and lubrication at the atomic scale. *Nature* **374**, 607–616 (1995)
16. Bhushan, B., Kulkarni, A.V., Koinkar, V.N., Boehm, M., Odoni, L., Martelet, C., Belin, M.: Microtribological characterization of self-assembled and Langmuir-Blodgett monolayers by atomic and friction force microscopy. *Langmuir* **11**, 3189–3198 (1995)
17. Bhushan, B., Kasai, T., Kulik, G., Barbieri, L., Hoffmann, P.: AFM study of perfluorosilane and alkylsilane self-assembled monolayers for anti-stiction in MEMS/NEMS. *Ultramicroscopy* **105**, 176–188 (2005)
18. Bhushan, B., Cichomski, M., Hoque, E., DeRose, J. A., Hoffmann, P., Mathieu, H.J.: Nanotribological characterization of perfluoroalkylphosphonate self-assembled monolayers deposited on aluminum-coated silicon substrates. *Microsyst. Technol.* **12**, 588–596 (2006)
19. Bhushan, B., Hansford, D., Lee, K.K.: Surface modification of silicon surfaces with vapor phase deposited ultrathin fluorosilane films for biomedical devices. *J. Vac. Sci. Technol. A* **24**, 1197–1202 (2006)
20. Callister, W.D.: Materials Science and Engineering, 4th edn. Wiley, New York (1997)
21. DeRose, J.A., Hoque, E., Bhushan, B., Mathieu, H.J.: Characterization of perfluorodecanote self-assembled monolayers on aluminum and comparison of stability with phosphonate and siloxy self-assembled monolayers. *Surf. Sci.* **602**, 1360–1367 (2008)
22. DeVecchio, D., Bhushan, B.: Localized surface elasticity measurements using an atomic force microscope. *Rev. Sci. Instrum.* **68**, 4498–4505 (1997)

23. Eustathopoulos, N., Nicholas, M., Drevet, B.: Wettability at High Temperature. Pergamon, Amsterdam (1999)
24. Hoque, E., DeRose, J.A., Hoffmann, P., Mathieu, H.J., Bhushan, B., Cichomski, M.: Phosphonate self-assembled monolayers on aluminum surfaces. *J. Chem. Phys.* **124**, 174710 (2006)
25. Hoque, E., DeRose, J.A., Kulik, G., Hoffmann, P., Mathieu, H.J., Bhushan, B.: Alkylphosphonate modified aluminum oxide surfaces. *J. Phys. Chem. B* **110**, 10855–10861 (2006)
26. Hoque, E., DeRose, J.A., Hoffmann, P., Bhushan, B., Mathieu, H.J.: Alkylperfluorosilane self-assembled monolayers on aluminum: a comparison with alkylphosphonate self-assembled monolayers. *J. Phys. Chem. C* **111**, 3956–3962 (2007)
27. Hoque, E., DeRose, J.A., Hoffmann, P., Bhushan, B., Mathieu, H.J.: Chemical stability of nonwetting, low adhesion self-assembled monolayer films formed by perfluoroalkylsilazation of copper. *J. Chem. Phys.* **126**, 114706 (2007)
28. Hoque, E., DeRose, J.A., Bhushan, B., Mathieu, H.J.: Self-assembled monolayers on aluminum and copper oxide surfaces: surface and interface characteristics, nanotribological properties, and chemical stability. In: Bhushan, B., Fuchs, H., Tomitori, M. (eds.) Applied Scanning Probe Methods Vol. IX – Characterization, pp. 235–281. Springer, Heidelberg (2008)
29. Hoque, E., DeRose, J.A., Bhushan, B., Hippis, K.W.: Low adhesion, non-wetting phosphonate self-assembled monolayer films formed on copper oxide surfaces. *Ultramicroscopy* **109**, 1015–1022 (2009)
30. Kasai, T., Bhushan, B., Kulik, G., Barbieri, L., Hoffmann, P.: Nanotribological study of perfluorosilane SAMs for anti-stiction and low wear. *J. Vac. Sci. Technol. B* **23**, 995–1003 (2005)
31. Lee, K.K., Bhushan, B., Hansford, D.: Nanotribological characterization of perfluoropolymer thin films for BioMEMS applications. *J. Vac. Sci. Technol. A* **23**, 804–810 (2005)
32. Lide, D.R.: CRC Handbook of Chemistry and Physics, 85th edn. CRC Press, Boca Raton (2004)
33. Liu, H., Bhushan, B.: Investigation of nanotribological properties of alkylthiol and biphenyl thiol self-assembled monolayers. *Ultramicroscopy* **91**, 185–202 (2002)
34. Liu, H., Bhushan, B.: Orientation and relocation of biphenyl thiol self-assembled monolayers. *Ultramicroscopy* **91**, 177–183 (2002)
35. Liu, H., Bhushan, B., Eck, W., Stadler, V.: Investigation of the adhesion, friction, and wear properties of biphenyl thiol self-assembled monolayers by atomic force microscopy. *J. Vac. Sci. Technol. A* **19**, 1234–1240 (2001)
36. Man, K.F., Stark, B.H., Ramesham, R.: A Resource Handbook for MEMS Reliability. Rev. A. JPL Press, Jet Propulsion Laboratory, California Institute of Technology, Pasadena (1998). See also, Man, K.F.: MEMS reliability for space applications by elimination of potential failure modes through testing and analysis. <http://www-rel.jpl.nasa.gov/Org/5053/atpo/products/Prod-map.html> (2002)
37. Ren, S., Yang, S., Zhao, Y., Yu, T., Xiao, X.: Preparation and characterization of ultrahydrophobic surface based on a stearic acid self-assembled monolayer over polyethyleneimine thin films. *Surf. Sci.* **546**, 64–74 (2003)
38. Tambe, N.S., Bhushan, B.: Nanotribological characterization of self assembled monolayers deposited on silicon and aluminum substrates. *Nanotechnology* **16**, 1549–1558 (2005)
39. Tanner, D.M., Smith, N.F., Irwin, L.W., et al.: MEMS Reliability: Infrastructure, Test Structure, Experiments, and Failure Modes. Sandia National Laboratories, Albuquerque (2000). SAND2000-0091
40. Tao, Z., Bhushan, B.: Degradation mechanisms and environmental effects on perfluoropolyether self assembled monolayers and diamond like carbon films. *Langmuir* **21**, 2391–2399 (2005)
41. Tao, Z., Bhushan, B.: Surface modification of AFM silicon probes for adhesion and wear reduction. *Tribol. Lett.* **21**, 1–16 (2006)
42. Wang, Y., Bhushan, B.: Boundary slip and nanobubble study in micro/nanofluidics with atomic force microscope. *Soft Matter* **6**, 29–66 (2010)

Self-Assembled Nanostructures

► Self-Assembly of Nanostructures

Self-Assembled Protein Layers

► Interfacial Investigation of Protein Films Using Acoustic Waves

Self-Assembly

► Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly

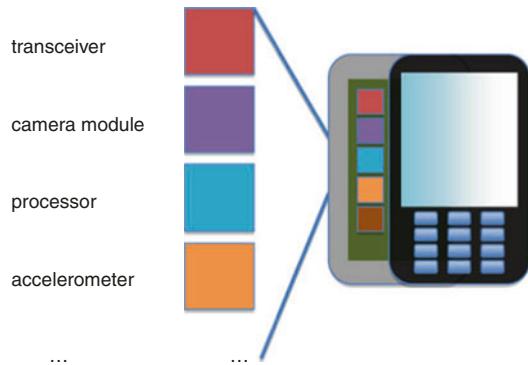
► Self-Assembly for Heterogeneous Integration of Microsystems

Self-Assembly for Heterogeneous Integration of Microsystems

J. H. Hoo¹, K. S. Park¹, R. Baskaran^{1,2} and K. F. Böhringer¹

¹Department of Electrical Engineering,
University of Washington, Seattle, WA, USA

²Components Research, Intel Corporation, Santa Clara, CA, USA



Synonyms

Heterogeneous integration; Self-assembly; Stochastic assembly

Definition

Self-assembly is defined as the autonomous organization of components into ordered patterns or structures without human intervention. In this entry, the scope of self-assembly will be restricted within the microelectronics field – typically, the process of transporting, aligning, and permanently adhering discrete components onto various substrates, achieving heterogeneous integration.

Introduction

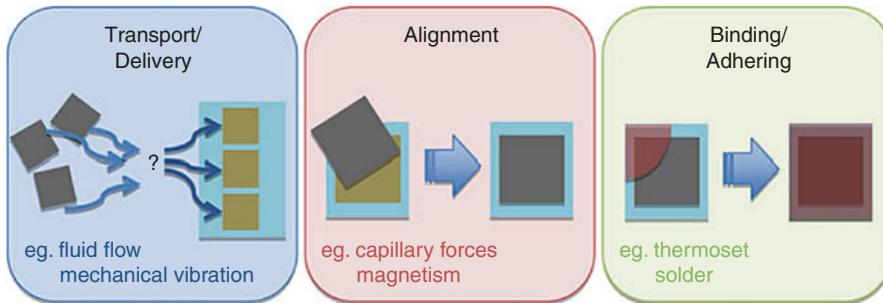
The packaging and integration of microscale modules, such as sensors, actuators and transceivers, and data processing, microfluidic, power management, electromechanical and optoelectronic devices into heterogeneous multifunctional systems is one of the big challenges in the field of microelectromechanical systems (MEMS). The fabrication processes and the material prerequisites for the components are predominantly incompatible, thereby eliminating the possibility of fabricating them all on a common substrate, as it would have been done in a monolithic approach. Consequently, in a heterogeneous approach, the components of a system are fabricated separately under respective optimal conditions before being assembled to construct a functional system (Fig. 1).

Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 1 A modern cell phone: an example of a multifunctional device put together by discrete components from incompatible fabrication processes

The prevailing method used for assembly and packaging in microelectronic manufacturing is robotic pick-and-place. This technique has proven accurate and reliable for large component scales, and its throughput is satisfying for consumer electronics applications [1]. However, pick-and-place is confronted with a trade-off between throughput and component placement accuracy. Furthermore, the method is serial, it requires closed-loop control, and it becomes more expensive as device dimensions shrink, and registration constraints become more stringent. Furthermore, stiction problems [2] set in for device sizes smaller than 300 μm [1].

Self-assembly, defined as *the autonomous organization of components into ordered patterns or structures without human intervention* [3], provides a promising alternative to pick-and-place machinery. Self-assembly is parallel in nature and can be applied to a wide range of sizes, from the millimeter to nanometer scales. Self-assembly processes can be engineered to avoid direct contact manipulation of components, thereby eliminating the issues with stiction.

Self-assembly is certainly not a suite of techniques exclusive to the microelectronic field. Self-assembling processes (also satisfying the stated definition) from the natural environment and reported in other scientific disciplines range from the noncovalent association of organic molecules to the formation of crystals from precipitates to



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 2 Self-assembly process can be separated into the transport, alignment, and permanent binding/adhering phases. During the transport phase,

components are brought to the immediate vicinity of binding/assembly sites; the alignment phase corrects the orientation of the component; the binding phase adheres the component onto the binding site permanently

DNA origami [4]. The scope of this entry will be restricted to the application of self-assembly within the microelectronics field – typically, the process of transporting, aligning, and permanently adhering discrete components onto various substrates (Fig. 2). It is important to note the stochastic nature of the transportation phase; it is impossible, and unnecessary, for a user of self-assembly to determine the specific component that gets delivered to a specific binding location. Table 1 tabulates some notable microscale self-assembly techniques in the field.

Self-assembly techniques can be classified by the underlying mechanisms that drive the transportation and alignment assembly steps (Table 1). Having introduced self-assembly as a possible alternative for prevailing methods of microcomponent assembly and packaging, metrics necessary to the evaluation and comparison of self-assembly techniques are also listed, including the dimensions of the discrete components to be assembled (lateral size, thickness, and aspect ratio), how closely can components be assembled on the substrate (packing density), if there is control over the orientation of the placement of the components (orientation specificity), and the average rate of completeness of the assembly processes (yield).

The successful application of self-assembly requires not only a match between the dimensions of the discrete components reported in Table 1 and that of the components one wishes to assemble. It is also important to consider the assembly

mechanisms, which are closely related to the environment in which the self-assembly takes place. For example, one should avoid any self-assembly techniques requiring magnetic force should the discrete components be sensitive to magnetic force. In the following sections of the entry, most of the assembly methodologies will be further examined.

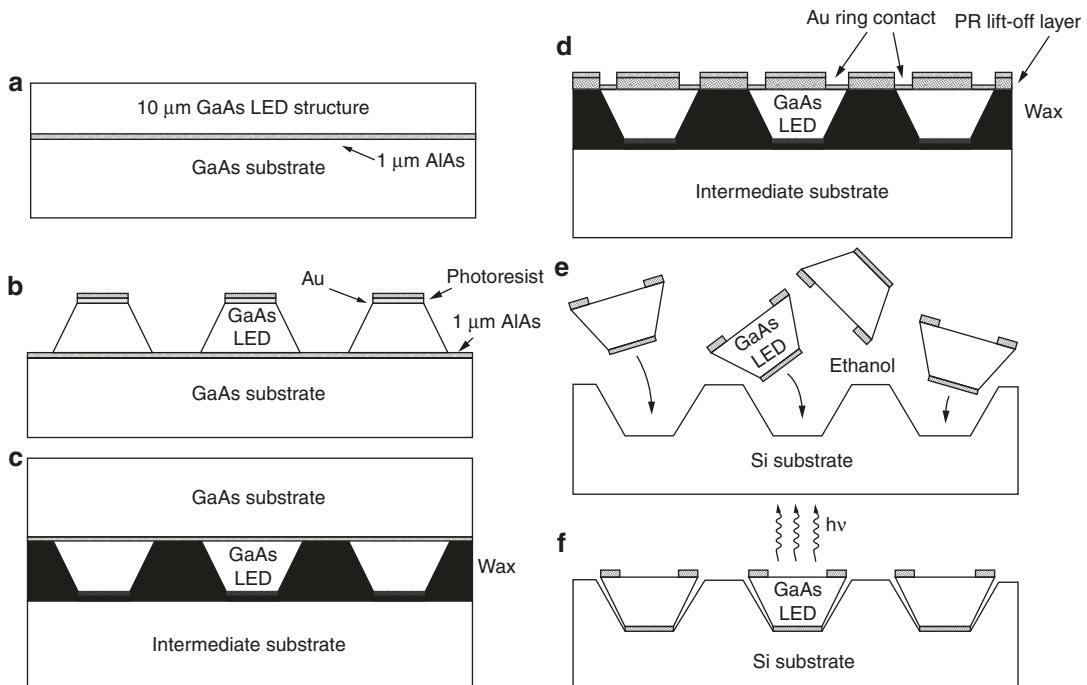
Fluidic Transport-Shape-Matching Alignment

Yeh et al. pioneered the work on self-assembly based on fluidic flow and shape matching [5]. Trapezoidal-shaped light-emitting diode (LED) devices were fabricated on GaAs substrates (Fig. 3) and released into ethanol or methanol, the carrier fluids of choice, which inhibits the oxidation and, therefore, degradation of the GaAs structures.

The LEDs were flowed over silicon host wafers fabricated with corresponding trapezoidal holes to capture them (Fig. 3). After the holes were (mostly) filled, the carrier fluid was evaporated, leaving the GaAs LEDs sitting in the holes, attached to the silicon surface only by van der Waals forces. The flowing of LEDs over the silicon substrate is identified as the transportation component of a self-assembly process. The alignment step was realized by shape matching between the trapezoidal LEDs and the corresponding silicon holes. The final procedure

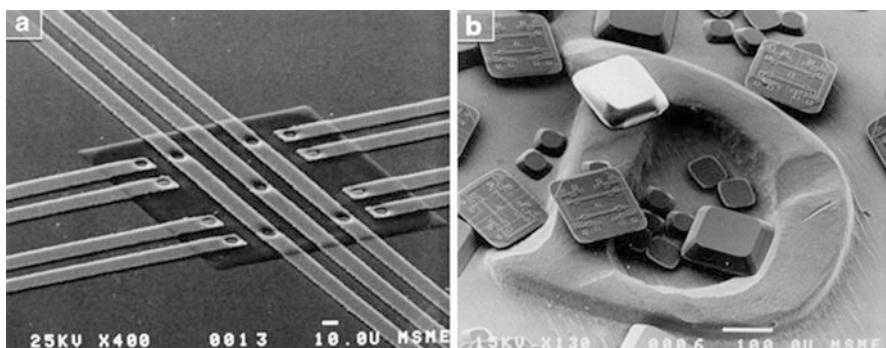
Self-Assembly for Heterogeneous Integration of Microsystems, Table 1 Survey of current self-assembly process. Components are square in shape except where marked by ^C, and ^R, which are circular and rectangular respectively. Part sizes of components are taken to be the length for square and rectangular parts, and the diameter for circular components

Researcher, institution	Assembly mechanism			Part size (lateral dim) (μm)	Thickness (μm)	Aspect ratio (lateral/thickness)	Packing density (in plane) (%)	Orientation specificity	Reported yield (%)
	Transport	Mechanical agitation	Alignment						
Yeh, Berkeley	✓		✓	18	9.9	1.8	—	No	>90
Parviz, UW	✓		✓	100 ^C	10–20	5–10	20	Yes	97
Srinivasan, Berkeley	✓		✓	150–400	15–50	8–10	16	No	100
Jacobs, UMN & Harvard	✓		✓	20	10	2	25	No	>98
Böhringer, UW	✓		✓	1,000–2,000	100	10–20	25	No	100
Böhringer, UW	✓		✓	2,000	100	20	40	Yes	100
Böhringer, UW	✓	✓		370	150	2.47	80	No	100
Böhringer, UW	✓	✓		400 ^R	200	2	35	No	100
Ramadan, IME	✓	✓	✓	1,000 ^C	350	2.86	6	No	97
Fischer, KTH	✓	✓	✓	35 ^C	350 ^L	—	—	No	>95



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 3 Schematic diagram of the fluid self-assembly process: (a) Molecular beam epitaxy (MBE) grown structure with 1 μm AlAs etch-stop layer, (b) trapezoidal GaAs mesa definition, (c) bonding to intermediate substrate with wax, (d) top-side ring contact

metallization, (e) solution containing the GaAs blocks dispensed over patterned Si substrate and (f) Si substrate with GaAs, light-emitting diodes integrated by fluidic self-assembly (Reprinted with permission from [5], © 1994 IEEE)



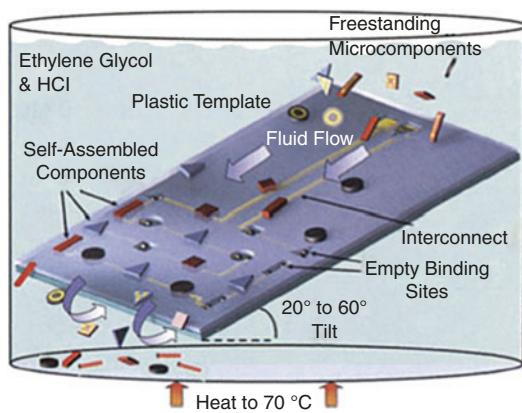
Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 4 Scanning electron microscope (SEM) picture of Yeh's devices: (a) a device assembled using fluidic self-assembly techniques – the metallic

connections are deposited over the device such that mechanical and electrical connection is established. (b) Devices placed on a US dime for scale (From [6], reprinted with permission from Alien Technology)

of binding was performed by the deposition of metallic connections between the devices and the host substrate, establishing electrical and mechanical connection, permanently adhering the chips and the substrate (Fig. 4). Yeh's self-assembly

technique was subsequently commercialized, and has been adapted to be a high volume manufacturing compatible process, specializing in the production of radio frequency identification (RFID) tags [6].

By combining Yeh's concept of shape matching under fluidic flow with capillary forces as drivers of self-assembly, Parviz et al. demonstrates the possibility to assemble multiple device types simultaneously onto a variety of substrates [1], including flexible plastics [7]. Figure 5 summarizes the features of Parviz's multi-device assembly process.



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 5 The heterogeneous self-assembly process. Microcomponents are introduced over a template submerged in a liquid medium and moved with the fluid flow. Self-assembly occurs as microcomponents first fall into complementarily shaped wells and then become bound by the capillary forces resultant from a molten alloy (Reprinted from [7], with permission from IOP)

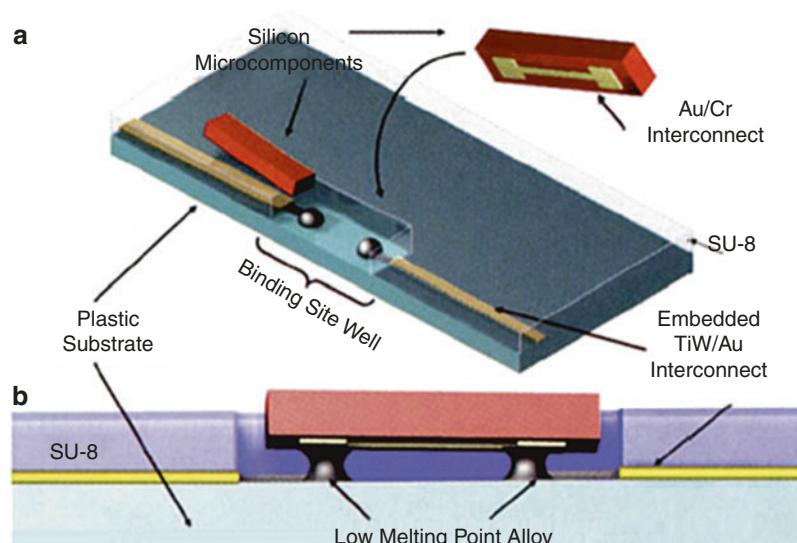
Self-Assembly for Heterogeneous Integration of Microsystems

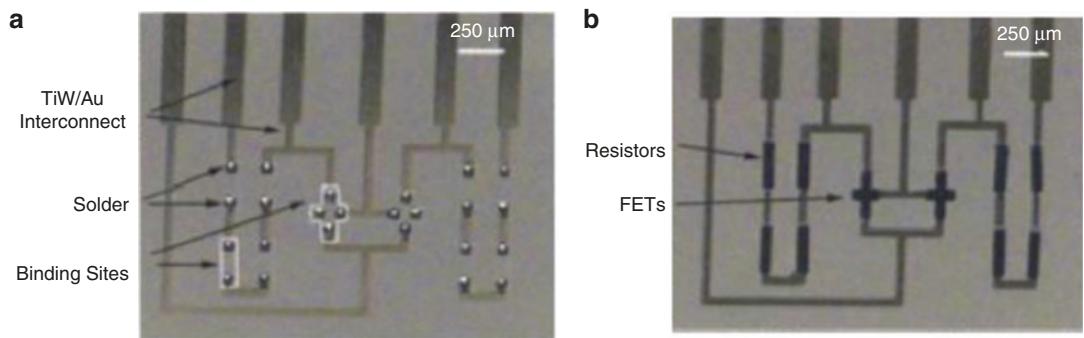
Fig. 6 Details of the self-assembly process for a single microcomponent. (a) The microcomponent approaches a binding site with complementary shape. (b) The microcomponent is held by capillary force resultant from molten-alloy-bridging the metal pads positioned on the microcomponent and on the template (Reprinted from [7], with permission from IOP)

Components of distinct shapes were flowed down the inclined surface of a template. Binding sites patterned on the template correspond to the shapes of components to be assembled. Component–binding site pairs are designed to only be compatible with each other, disallowing unintended assembly. The system is also designed to allow binding sites to only accept components when it happens to approach with the correct surface facing the binding site (Fig. 6).

The assembly process (Figs. 5 and 6) was maintained above the melting temperature of the low-melting-point alloy solder deposited at the contact pads of the template, within the binding sites. When the metal contact pads on a single microcomponent comes into contact with the molten solder, the component will be pulled to the template as the molten solder wets the contact pad on the chip. Conversely, if a component approaches the binding site with the incorrect side, capillary forces will not be exerted on the microcomponent, as molten solder will not wet on silicon surfaces. Excess components collected at the bottom of the container will be reused repeatedly until assembly is complete (Fig. 5).

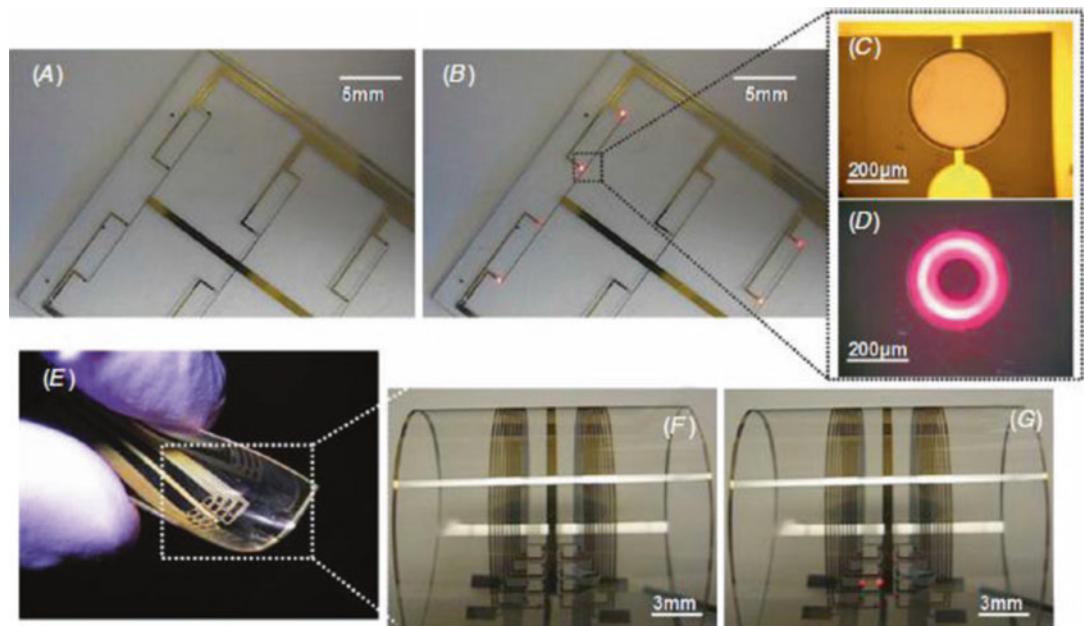
When assembly was observed to be complete, the temperature of the self-assembly setup was lowered to room temperature, thereby solidifying the molten solder, establishing electrical and





Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 7 Self-assembly of field effect transistors (FETs). (a) Optical microscopic image of a plastic substrate with empty binding sites (two outlined

with white lines for clarity); (b) The template after completion of the self-assembly process showing the position of FETs and diffusion resistors (Reprinted from [7], with permission from IOP)



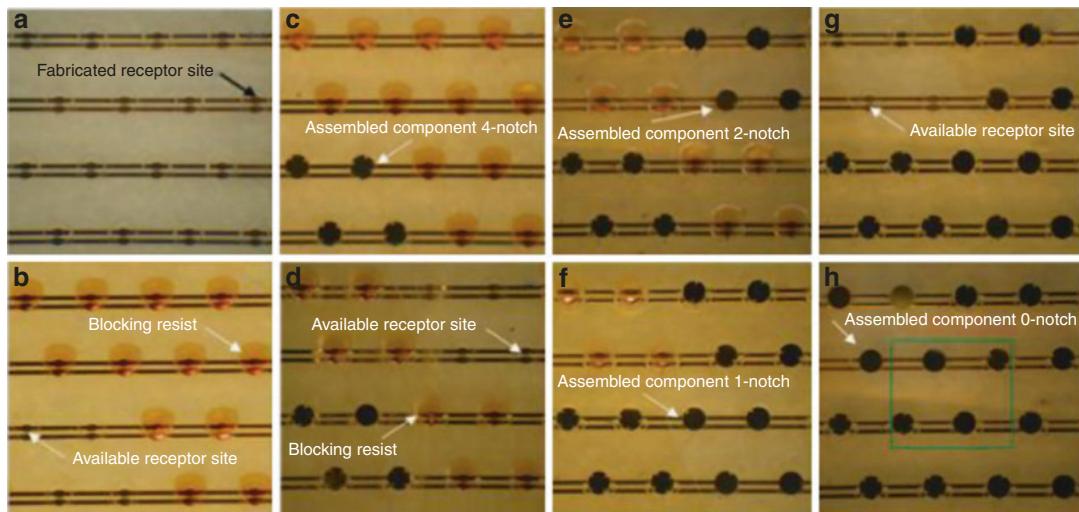
Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 8 (a) A glass template after the conclusion of the self-assembly process. The micro (μ)-LEDs are assembled on the template. (b) 6 μ -LEDs are turned ON on the same glass template via the application of a 5 V bias. (c), (d) Close-up images of 1 μ -LED turned OFF

and ON, on the template under an optical microscope; (e) A flexible plastic template after the completion of the self-assembly process; (f, g) The same plastic template, bent over itself, with μ -LEDs OFF and then ON upon the application of a 4 V bias (Reprinted from [7], with permission from IOP)

mechanical connection between the template and the discrete components. While the fluidic flow transportation mechanism is similar to Yeh et al., the alignment of microcomponents is achieved through shape-matching and solder capillary action. Finally, binding of the discrete

components was achieved through the solidification of the molten solder bumps. Final products are shown in Figs. 7 and 8.

Parviz et al. added programmability to their fluidic shape-matching self-assembly process with a technique that involves the use of



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 9 Optical microscope images of a sequence of events during the self-assembly process of four different microcomponent types. (a) Fabricated receptor site wells and C-shaped traps on a plastic template prior to self-assembly; (b) Positive photoresist was patterned on the template to block all the receptor sites, except the four receptor sites located on the *bottom left*. (c) The first type of components (circular 4-notch) was assembled within the 2×2 parallelogram pattern of available receptor sites. (d) A thin layer of photoresist was coated on the whole

template to fix the assembled microcomponents. The blocking resist was then removed from the four top right receptor sites designated for the second type of microcomponents. (e) The second type of components (circular 2-notch) was assembled in available receptor sites. (f) The third type of components (circular 1-notch) was assembled in the *bottom left corner* of the image in a similar fashion. (g) The resist was removed from the remaining four receptor sites on the *top left*. (h) Final type of components (circular 0-notch) was self-assembled (Reprinted from [8], with permission from MRS)

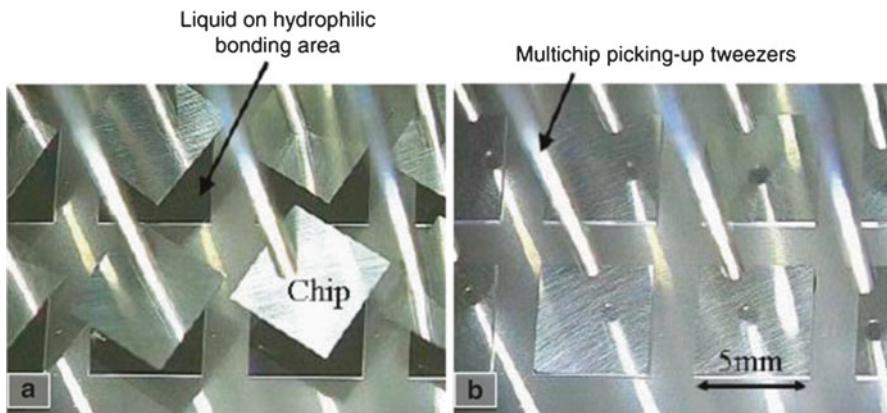
photosensitive materials to control the access of binding sites to microcomponents being flowed across the template [8]. Programmable binding locations were created by depositing photoresist within receptor sites that uses the principle of shape matching to capture microcomponents that are flowed over them. To activate a group of receptor sites, the region of interest is selectively exposed to ultraviolet radiation to remove the blocking photoresist material, thereby availing said sites to capture incoming components. By successive steps of site activation and component flowing, different components can be assembled onto a single type of receptor sites (Fig. 9).

It is instructive, at this point, to definite “programmability” in the context of micro self-assembly as *the ability to direct different reliably reproducible outcomes with the same set of components in a self-assembly process.*

Programmability, which will be revisited in other works, will add to the appeal of self-assembly in industry.

Fluidic Transport–Capillary Forces Alignment

Capillary forces dominate inertial forces and gravity as one approaches submillimeter scales, given that capillarity scales with length, while gravity scales with length cubed. There are many studies that concentrate on the alignment properties of capillary forces instead of an entire self-assembly process as defined in Fig. 2. Notably, work from Koyanagi et al. focuses on the aligning and bonding of silicon chips on micro-fabricated silicon pedestals with corresponding dimensions [9] (Fig. 10).



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 10 Self-alignment process from Koyanagi et al. (a) Introduction of chips using a multi-chip

vacuum pick system. (b) Chips self-aligned after placement on bonding areas (Reprinted with permission from [9], © 2008 IEEE)

Droplets of an aqueous solution of hydrofluoric acid (HF) were placed on the pedestals before introducing silicon chips with a multi-chip vacuum pick system. The faces of the chips and the pedestals were treated to be hydrophilic so that the droplet underneath the chips will wet both surfaces and self-align the chips to the pedestals underneath to minimize the interfacial energies. Alignment will improve as the HF solution evaporates, and eventually dissipates entirely, while bonding chips to pedestals permanently with strong SiO₂–SiO₂ bonds.

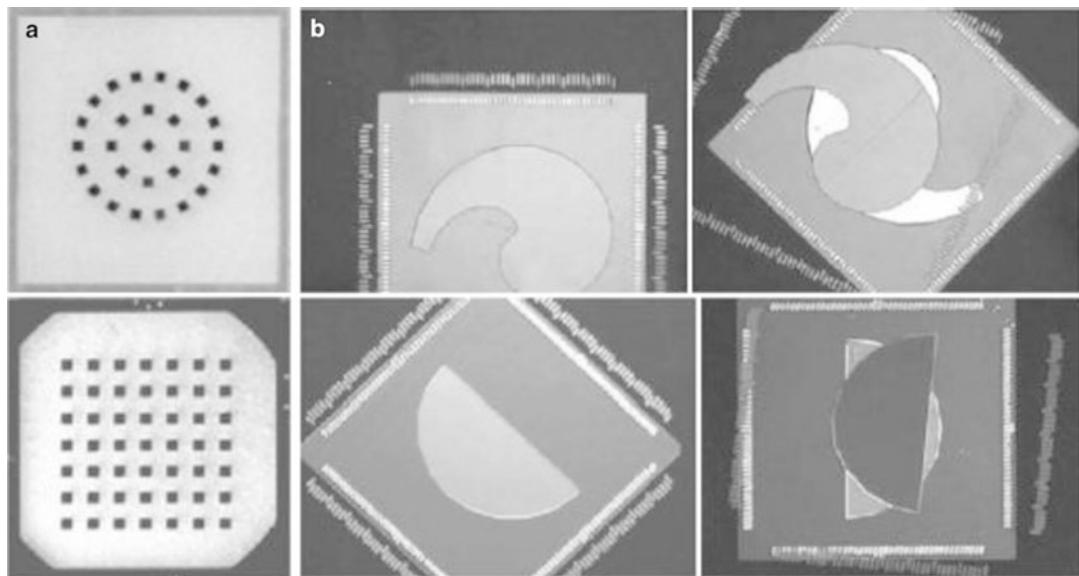
There are many other works titled “self-assembly” that deal exclusively with alignment rather than the entire self-assembly process which were presented in Fig. 2. “Self-alignment” is probably a more appropriate title for such studies, drawing the distinction between investigations focused on alignment and works that also include a stochastic transportation mechanism that brings discrete components to the vicinity of the intended assemblies.

Srinivasan et al. demonstrated a complete self-assembly technique with fluidic component transportation and capillary force capture and alignment [10]. They patterned a substrate with an array of hydrophobic self-assembled monolayer (SAM)-coated gold-binding sites. By inserting the substrate into water through a film of

hydrophobic adhesive floating on the water surface, the hydrophobic binding sites were coated with the adhesive (Fig. 11).

Microparts fabricated from silicon-on-insulator (SOI) wafers were then introduced through a pipette toward the substrate in water. When the hydrophobic pattern on the microparts came into contact with the adhesive, the parts were pulled to the hydrophobic binding sites and self-alignments occurred spontaneously. Finally, when all sites were occupied, the adhesive was polymerized by heat (thermoset) or UV radiation, depending on the type of adhesive used, bonding the chips to the substrate permanently. Binding sites of shapes with in-plane rotational symmetries such as squares gave alignment yields of up to 100 %, with translation and rotational misalignments of less than 0.2 μm and 0.3°, respectively. Component–binding site pairs without in-plane rotational symmetries (aimed at specific in-plane alignment) such as semicircles and commas gave alignments yields of approximately 30–40 %.

Jacobs et al. developed a similar self-assembly process that uses low-temperature melting solder to mount LED arrays on flexible cylindrical templates [11]. The assembly template was patterned with copper squares, and a simple dip-coating process applied molten solder on these copper squares, given that copper has good wetting



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 11 Optical micrographs of capillary-driven self-assembly of flat microparts: (a) Square parts on quartz substrates in *ring* and grid

configurations and (b) correct and wrong alignments for *semicircle*- and *comma-shaped* binding sites (Reprinted with permission from [10], © 2001 IEEE)

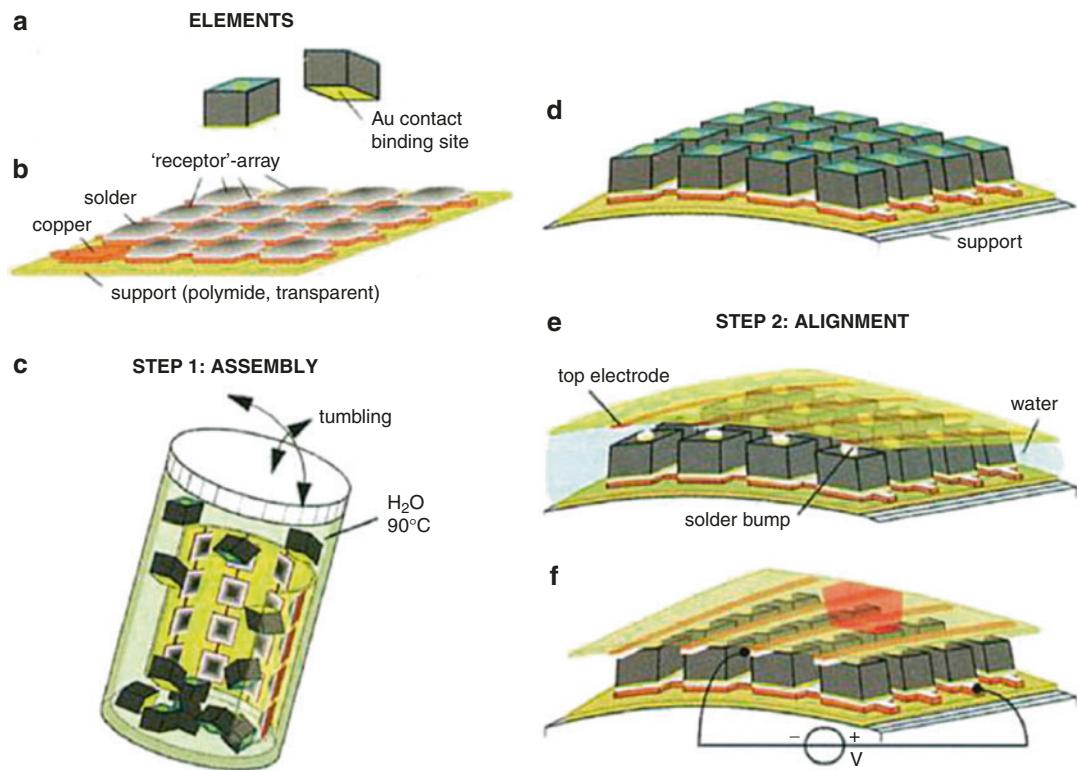
characteristics for solders. The dip-coating process was performed in an acidic aqueous environment to prevent oxidation of the solder surface.

Hundreds of LEDs and a flexible assembly template were placed inside a vial; the vial was filled with water and heated to a temperature above the melting point of the solder. The LEDs were tumbled inside the vial, allowing for the LEDs to come into contact with the molten solder, be pulled to a copper square by the solder, and self-aligned (Fig. 12c). The cathode and anode of the LEDs were distinguished by a small round contact and a larger square contact (the size of the copper squares on the template), respectively, on two faces of the discrete LEDs (Fig. 12a). The agitation intensity of the tumbling process was tuned to dissociate assemblies between the solder and the smaller round contact while not severe enough to affect LEDs that were caught with their square contact face. This gives the assembly process side-selectively.

A permanent bond between each LED and the template was established when the template was allowed to cool, solidifying the solder. To make the LED array functional, a transparent film

patterned with electrical circuitry was aligned and bonded to the assembled array, establishing the electrical connections to the cathodes of the LEDs. By activating the LEDs (Fig. 13), the defect rate was visually determined to be approximately 2 %.

Jacobs et al. also worked on self-assembly based on the delivery of components on a fluid-fluid interface [12]. Components were fabricated out of silicon and SU-8, with one face coated with gold. The gold surfaces were treated with a mercaptoundecanoic acid (MUA) SAM to render them hydrophilic. The silicon faces were treated to be hydrophobic using 3-glycidoxypropyltrimethoxysilane (GPTMS), creating an orientation preference for components when they are introduced at a silicone oil–water fluid-fluid interface (SU-8 surface are hydrophobic and thus required no further surface modification). By mildly agitating the system, components at the silicone oil–water interface will predominantly orient themselves such that the hydrophilic gold surface will face toward water, and the hydrophobic silicon/SU-8 will face the direction of the silicone oil.



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 12 Procedure used to assemble a functional cylindrical display. (a) Top and bottom views of an LED segment that has two contacts: a small circular gold contact (cathode) on the front, and a large square gold contact (anode) on the back. (b) Array of solder-coated copper squares supported on a flexible substrate; these squares are of the same size as the anodes of the LEDs and act as receptors for the LEDs during the self-assembly. (c) The components are tumbled in a vial at a temperature

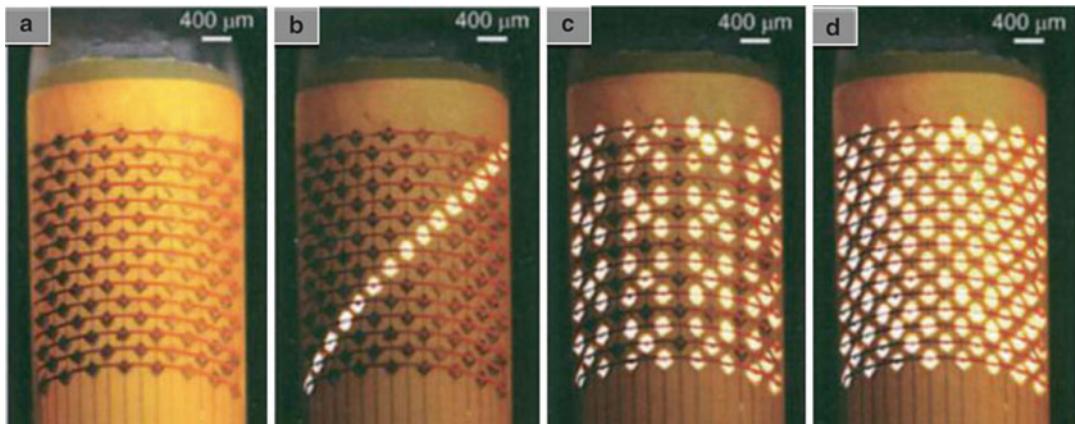
above the melting point of the solder. (d) Two-dimensional array of the assembled LEDs. (e) Alignment of a top electrode: The copper wires of the top electrode and the cathodic contacts on the front of the LEDs were first dip coated with solder. The array of wires is prealigned with the array of cathodic solder bumps. At a temperature above the melting point, electrical connections form and the structure self-aligns. (f) Test of the self-assembled display-prototype (From [11], reprinted with permission from AAAS)

Two types of substrates have been used: silicon and flexible propylene terephthalate (PET). Copper contact pads, corresponding to the sizes of the components, were patterned on the surface. A dip-coating process was used to deposit low-melting-temperature solder onto the copper pads.

The transfer of the components is shown in Fig. 14. A container containing silicone oil, water, and components on the interface between the two fluids was first heated to a temperature beyond the melting point of the solder. Substrates were pulled upward through the liquid-liquid interface, allowing the components to come into contact with the molten solder. Upon contact, the molten solder will pull components in,

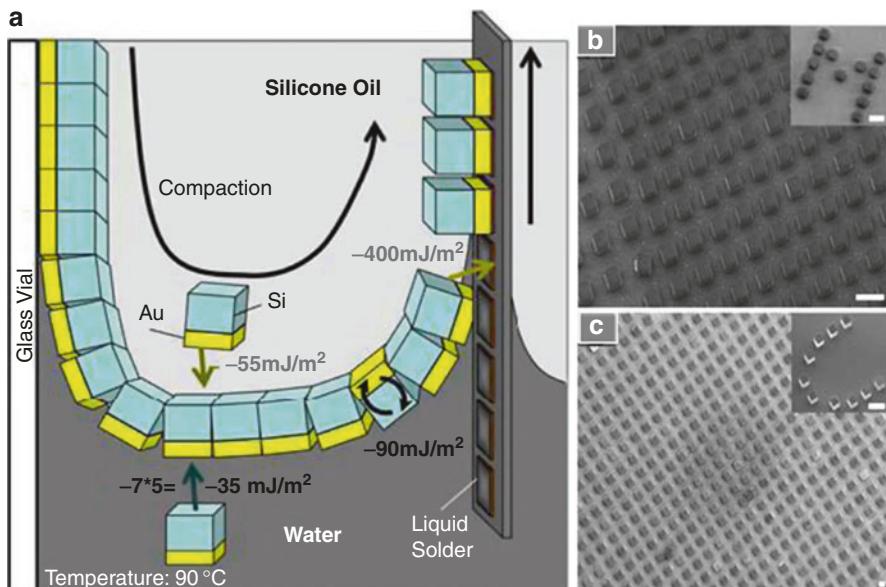
transferring the components onto the substrates. While it is possible to deliver components to over 90 % of the copper pads on a substrate in a single pass, multiple passes were required to achieve full coverage. Using this method, chips $20 \times 20 \times 10$ and $60 \times 60 \times 20 \mu\text{m}^3$ in size were assembled onto rigid and flexible substrates at reported speeds of up to 62,500 chips in 45 s.

Böhringer et al. studied the assembly of thin microcomponents in a fluidic environment, delivering components at the air-water interface [13]. Components ($1,000 \times 1,000 \times 100 \mu\text{m}^3$) with one face coated in gold are fabricated in silicon. Gold-coated silicon assembly templates were prepared with square openings, exposing



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 13 (a) Photograph of the display after self-alignment of the top electrode; (b–d) Photographs of the operating display after the alignment of the top electrode. The display contains 113 LEDs that are

assembled in an interleaved fully addressable array with eight columns of eight receptors interleaved with seven columns of seven on the *bottom* electrode that connect to 15 rows of crossing copper wires on the *top* electrode (From [11], reprinted with permission from AAAS)



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 14 (a) Procedure of surface tension-directed self-assembly at a liquid-liquid-solid interface. SEM of (b) SU-8 (20 μm side length) and (c)

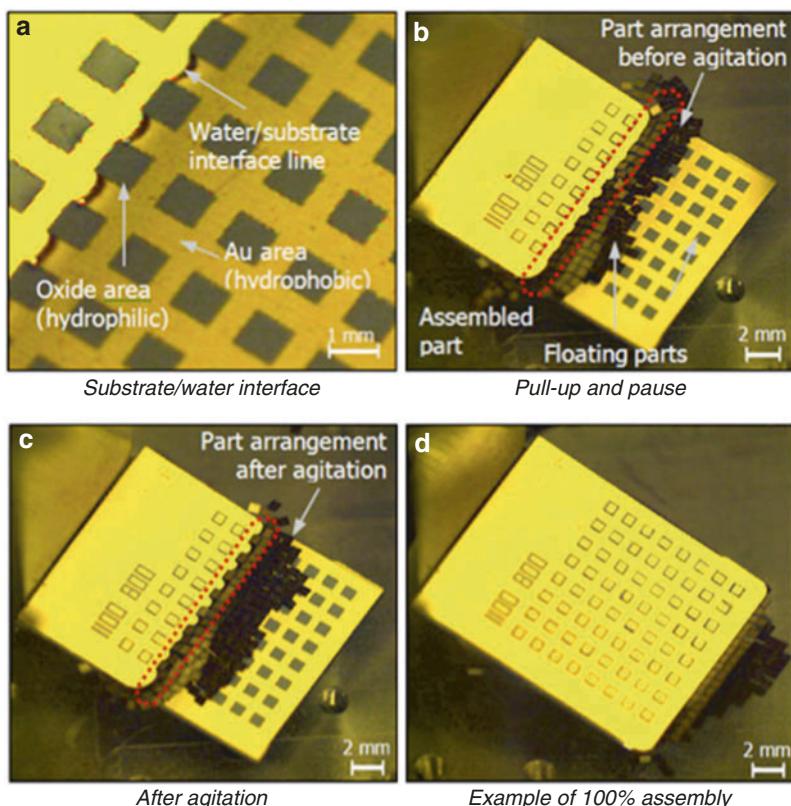
Si chiplets (20 μm and 60 μm side length) assembling in regular arrays and arbitrary text patterns (*insets*) (Reprinted from [12], © 2010 National Academy of Sciences, USA)

silicon, that were designed to be slightly larger than the lateral dimensions of the components. All gold surfaces were then treated to be hydrophobic using dodecanethiol SAM.

Components were introduced onto a water surface. By applying controlled agitation, the components will preferentially orient themselves with their gold faces upward due to the surface

**Self-Assembly
for Heterogeneous
Integration
of Microsystems,**

Fig. 15 Parts rearrangement by agitation. Darker squares are hydrophilic silicon surfaces



treatment. The assembly template was then pulled up, at an angle, through the air-water interface, in the immediate proximity of the floating components.

When the level of the assembly template reaches the top edges of a row of assembly sites (Fig. 15a), controlled agitation was applied to cause the components to arrange themselves such that the edge of a single part was pinned to the edge of a single site (Fig. 15b, c). At this point, the assembly template was pulled up, allowing an entire row of parts to be transferred and self-aligned to the template. By repeating this pull-up, pause and agitate cycle, entire assembly templates can be perfectly filled. The most significant aspect of this work lies in the aspect ratio of the components being self-assembled; no other self-assembly technique has been demonstrated to work well with very thin components.

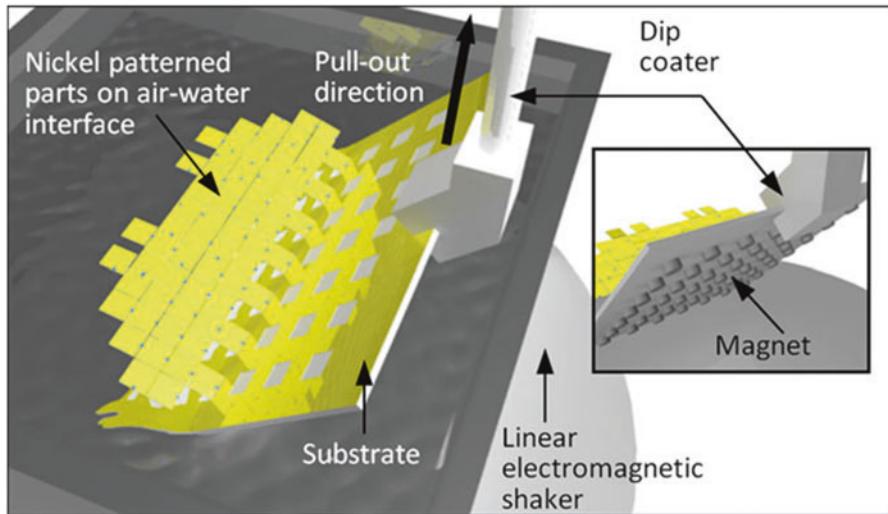
The programmability of this process was demonstrated by the ability to assemble only on

alternate rows of binding sites. After a single (first) row of parts have been transferred onto the assembly substrate, one can avoid assembling on the subsequent row by agitating the water surface while the substrate is extracted through the air-water interface, bypassing the second row and stopping at the assembly position of the third. Using similar methods, arbitrary row assemblies can be reproduced reliably.

S

Fluidic Transport–Capillary/Magnetic Forces Alignment

By depositing nickel squares near the edges of chips, and positioning a magnet underneath each binding site underneath the assembly substrate, Böhringer et al. extended the previous fluidic transport–capillary forces alignment self-assembly method to include orientation specificity (Fig. 16).



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 16 The experimental setup consists of a water container, linear electromagnetic actuator,

dip coater, assembly substrate, magnets, and test parts floating at the air-water interface

Similar to the previous method, controlled agitation of the water surface ensures one-to-one addressing of parts to binding sites. During this agitation, the interaction of the magnetic forces between the magnets underneath each binding site and the ferromagnetic nickel squares ensures also that chips approach the binding sites with the nickel-patterned edge, allowing for orientation specificity previously unachievable (Fig. 17).

consisting of a shape-matching step and a capillary force–driven alignment step [14–16] for square silicon parts as well as surface mount technology (SMT) passive devices.

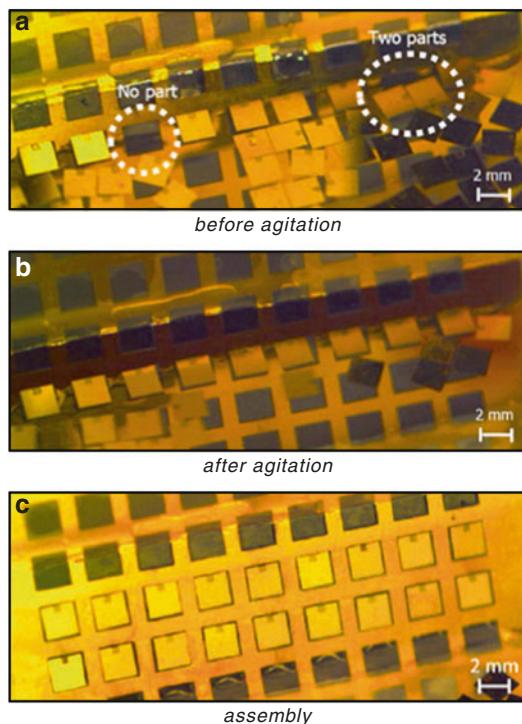
The latest iteration in the technique for the assembly of silicon chips is shown in Fig. 18. A silicon fabricated assembly template was first aligned on top of a transfer substrate. The transfer substrate was fabricated by first depositing gold onto a silicon wafer. Square openings (silicon) corresponding to the size of the microparts were subsequently developed, and the gold surfaces were treated to be hydrophobic with a SAM coating. Each aperture in the assembly template addresses a single silicon opening underneath (Fig. 18h), and can only contain a single discrete component standing on its side.

Square-shaped silicon microcomponents with one gold-coated (hydrophobic SAM-treated) side were delivered to every aperture on the assembly template. When the delivery was completed, the assembly template was removed, and steam was introduced to the system to form water droplets on the hydrophilic silicon surfaces. When the droplets become large enough, they will exert enough

Mechanical Agitation Transport–Capillary Forces Alignment

For discrete components and/or substrates that are adverse to fluidic environments, there are also self-assembly processes that are conducted in dry or “semidry” [14–16] conditions. Dry environment self-assembly processes predominantly require the use of mechanical vibration/agitation to distribute the microcomponents.

Böhringer et al. devised a suite of self-assembly methods to be performed in an air environment that uses a two-step assembly process



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 17 Assembly process of $2,000 \times 2,000 \times 100 \mu\text{m}^3$ components on the assembly substrate: (a) A binding site does not yet have a corresponding component, and another has two test parts in its vicinity. (b) After applying specific Faraday waves using agitation, every binding site gets a single component. (c) Two rows of components assembled onto the assembly substrate

force to pull the silicon face of the microparts flat onto the transfer substrate, and self-align the components to the squares. The aligned microcomponents can then be transferred onto a destination substrate to be permanently mounted. Similar techniques have been tested with parts ranging from $370 \times 370 \times 150 \mu\text{m}^3$ to $790 \times 790 \times 330 \mu\text{m}^3$ in size.

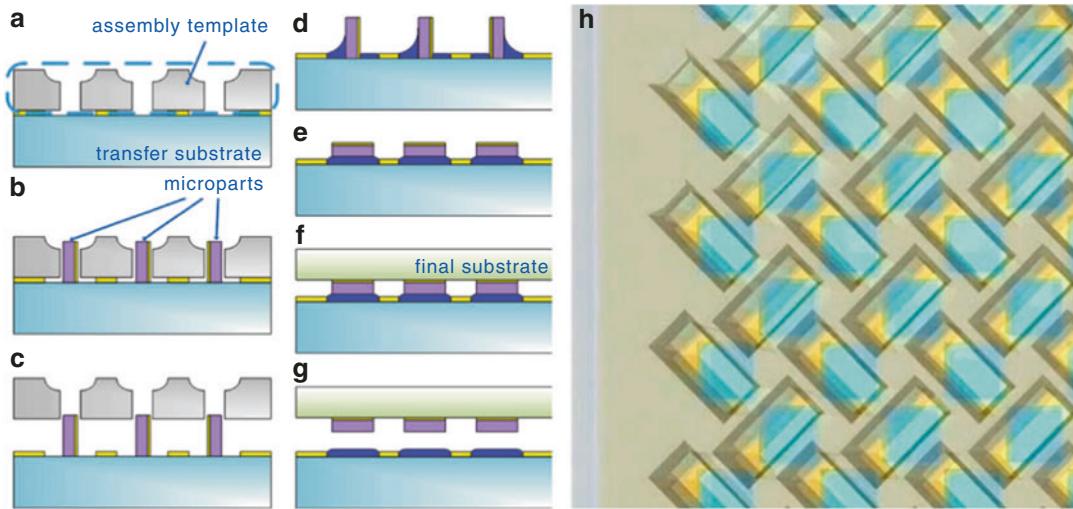
Böhringer et al. also developed a system of control for the precise manipulation of microcomponents on a vibrating surface [15]. Parts can be made to jump or walk (where parts seem to crawl across the assembly template) in predictable directions and speeds. Using these controllable part-motion modes, the

transportation phase of the self-assembly process can be programmed to deliver components to arbitrary regions of apertures or even empty unfilled apertures. In the case of a programmed “feedback-driven” delivery process [15] (Fig. 19), visual feedback is used to identify empty apertures and direct excess components toward them, guaranteeing the complete delivery of components.

Finally, Böhringer et al. also adapted the template-based methodology (complete with the feedback-driven assembly process) to the assembly of surface mount technology (SMT) thin-film resistors and monolithic ceramic capacitors of the 01005 standard ($0.016'' \times 0.008''$, 0.4×0.2 mm) [16]. In this instance, two templates were used, one stacked on top of another. Template 1 (Fig. 20) ensures that only one component will be allowed into template 2, and template 2 guides the in-plane orientation of the component. During the solder reflow process, the molten solder wets the metallized contact-ends of the SMT components and aligns them to the contact pads on the target substrate. At the end of the solder reflow, mechanical and electrical bonds between components and substrate were achieved.

Mechanical Agitation/Magnetic Forces Transport–Shape-Matching Alignment

A dry environment self-assembly technique combining transportation via mechanical vibration and magnetic attractive force with shape-matching alignment was devised by Ramadan et al. [17] (Fig. 21). A host substrate, featuring physical recesses, was superimposed to a master array comprising embedded NdFeB magnets, whose position exactly matched that of the recesses. Target chips, bearing a CoNiP soft magnetic film on their nonfunctional side, were distributed over the substrate, which was mechanically vibrated and tilted. This motion causes chips to be distributed across the entire substrate, until being collected in an enclosure at the lower end of the tilt.



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 18 Full assembly process flow (a–g): (a) Mount assembly template on a gold-patterned transfer substrate. (b) Assemble parts. (c) Remove template. (d) Apply moisture. (e) Vibrate the setup gently to have the parts fall on hydrophilic side and self-align. (f)

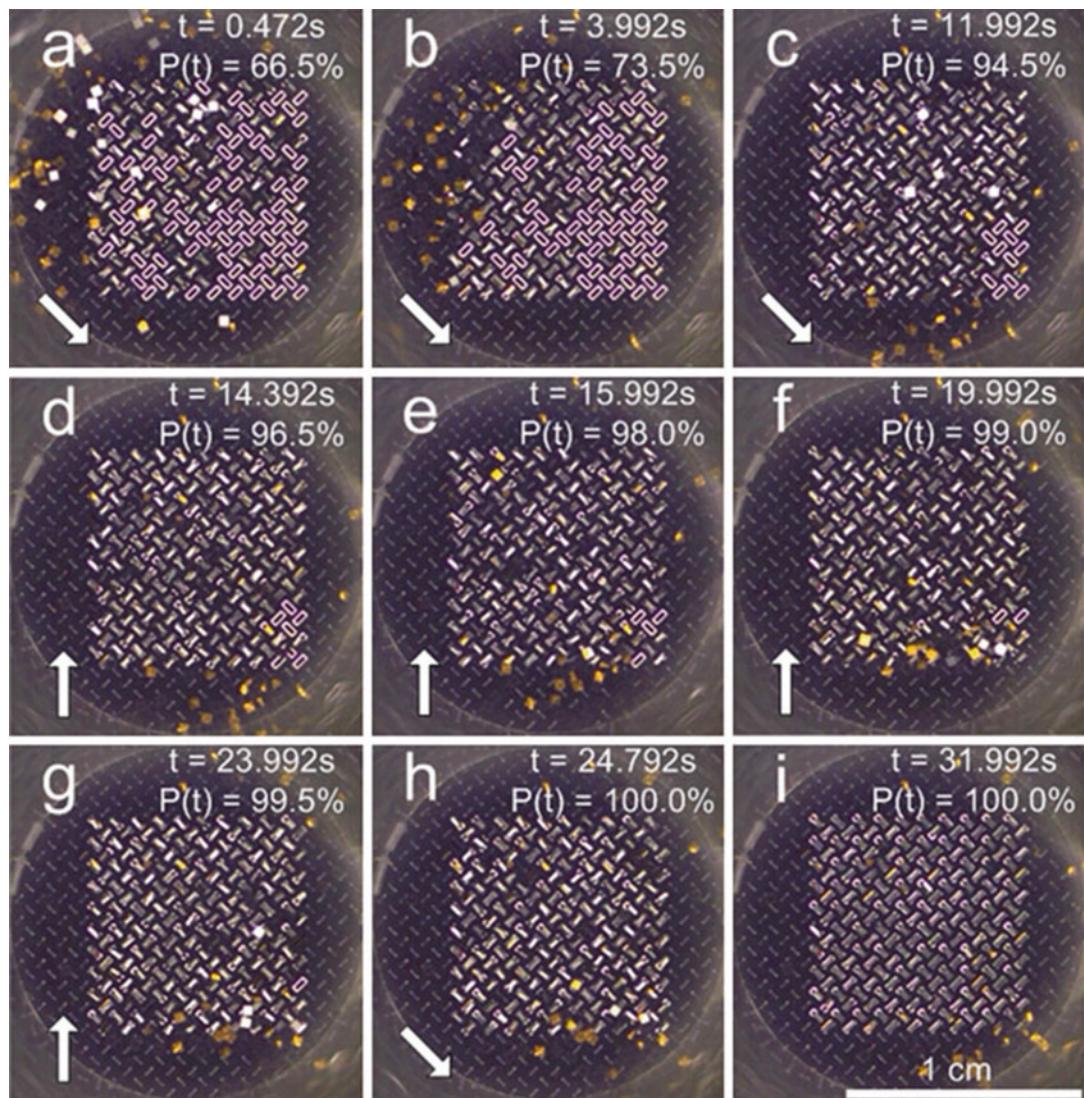
Attach parts to destination substrate/device. (g) Complete the process by removing the assembled substrate/device from transfer template. (h) Top-down view of assembly template placed over transfer substrate (Reprinted with permission from [15], © 2009 IEEE)

The amplitude of the substrate vibration and the thickness of chips were tailored so that the only stable configuration possible for the chips to be assembled is to have the functional side facing upward. The in-plane alignment of the chips was controlled by shape matching between the chips and the recesses of the host substrate. After assembly, the master array could be removed and reutilized while the assembled substrate could be further processed. An assembly of 2,500 chips in 5 min with 97 % yield was reported for this technique.

Fischer et al. proposed a scheme to implement through silicon vias (TSVs) by using magnetic force [18]. High-aspect-ratio holes are first etched on a silicon substrate using deep reactive ion etching (DRIE). Commercially available nickel wires (length: 350 μm , diameter: 35 μm) were deposited on the surface of the silicon substrate. A magnet was introduced at the bottom of the substrate, causing the wires to stand upright. By moving the magnet laterally, wires were dragged along the surface (standing), and pulled into the holes (Fig. 22).

After the wires have been assembled into the holes (Fig. 23), the excess space within a wire-occupied hole was filled by a manual application of the thermosetting polymer bisbenzocyclobutene (BCB). The BCB was then cured, permanently adhering the nickel wires within the silicon substrate. Subsequent steps involve steps to expose the ends of the wires on two sides of the silicon, and depositing metal contact lines per specific application. The filling rate of a 30 by 30 array of via holes with a pitch of 120 μm was reported at about 80 % in 20 s, and near complete (>95 %) in about 2 min.

TSVs are vertical electrical connections passing through die layers. They are important for heterogeneous integration solutions involving the stacking of dies. While seemingly out of place with previously mentioned assembly methodologies for chips/dies and passive devices, it is significant to show that there is a viable self-assembly solution for TSV placements, demonstrating that there are self-assembly solutions for every aspect of the device assembly process.



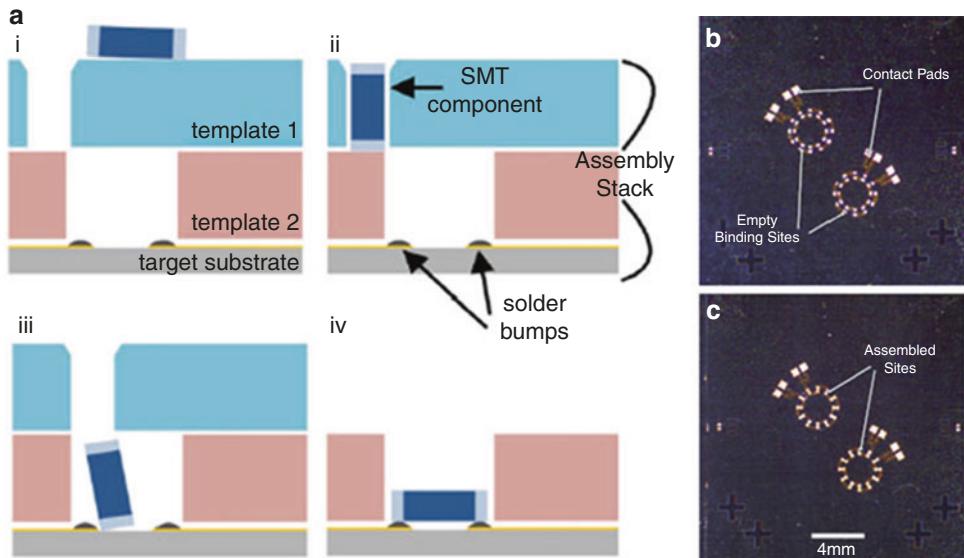
Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 19 Stages of a feedback-driven assembly: (a) shows the instance when jumping mode (with a *top left corner* bias) is switched to a walking mode with part-motion direction as indicated by the *white arrow* in the *box*. (b) and (c) show the progressive filling of the entire assembly area. *Right* after (c), when assembly-percentage, $P(t)$, has plateaued, a walking mode in the

upward direction as seen in (d–g) is activated, moving excess parts below the assembly area to the empty sites. After achieving 100 % assembly in (h), a walking mode directed toward the lower-right corner is activated, moving excess parts away from the assembly area, finishing the process in (i); note: size-scale provided in image (i) (Reprinted with permission from [15], © 2009 IEEE)

Conclusion

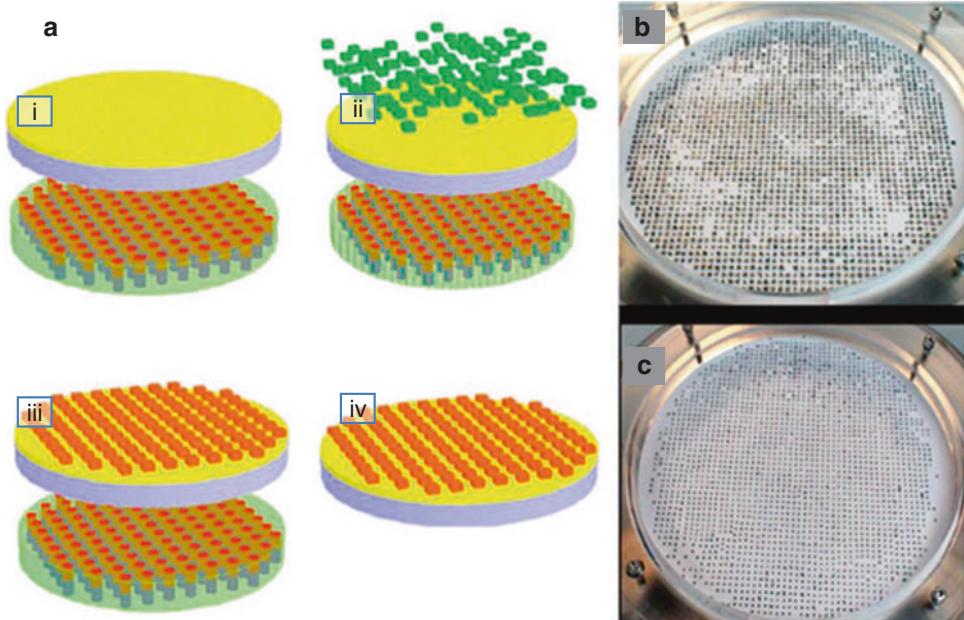
Microscale self-assembly has been presented as a suite of techniques that can enable the advancement of the electronics industry through

heterogeneous integration. A selection of self-assembly methodologies that are representative of techniques employing various modes of transportation and alignment, and the associated environments in which they are performed have been



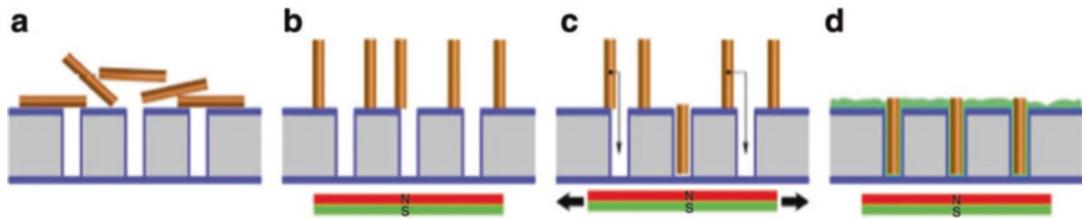
Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 20 (a) Assembly process: (i) Walking mode [3] component delivery performed on assembly stack in Configuration 1; (ii) a single component is captured by template 1 near the binding location; (iii) Configuration 2 – component is allowed to drop into template 2; (iv) slightly agitating the system aligns the

component to the orientation of the aperture on top of the binding location on the target substrate. After step (iv), solder reflow is performed to bond the component to the target substrate mechanically and electrically. Assembly of SMT passive components onto a circular test device: (b) before assembly; (c) after assembly (Reprinted with permission from [16], © 2009 IEEE)



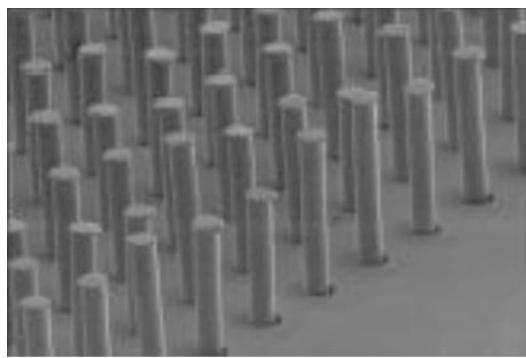
Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 21 Assembly process (a): (i) Host substrate aligned to the master magnetic array, (ii) chip seeding, (iii) applying vibration, and (iv) chip final

alignment and master magnetic array removal. Photo of 1 mm² assembly (b) 2 mins and (c) 5 mins into the vibration process (Reprinted with permission from [17], © 2007 American Institute of Physics)



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 22 Nickel wire assembly process: (a) nickel wires are introduced to a substrate etched with corresponding holes, (b) a magnet is placed under the substrate, causing the nickel wires to stand upright due to magnetic forces, (c) the magnet is moved laterally under

the substrate, moving wires to unoccupied holes and trapping them, (d) BCB is applied over the assembled wires. Processing steps after (d) include curing the BCB, polishing the substrate to expose the ends of the nickel wires, and depositing metal lines for electrical connection (Reprinted with permission from [18], © 2011 IEEE)



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 23 SEM image of a 30 × 30 array with a pitch of 120 μm of nickel wires placed in via holes prior to the filling of BCB (Reprinted with permission from [18], © 2011 IEEE)

introduced. By identifying self-assembly techniques with suitable discrete component dimensions as well as carefully considering the transport and alignment mechanism and assembly environment, one can harness the parallel and scalable nature of self-assembly, thereby streamlining the existing assembly processes or enabling revolutionary pathways.

References

- Morris, C.J., Stauth, S.A., Parviz, B.A.: Self-assembly for microscale and nanoscale packaging: steps toward self-packaging. *IEEE Trans. Adv. Packag.* **28**, 600–611 (2005)
- Fearing, R.S.: Survey of sticking effects for micro parts handling. Presented at the proceedings of the international conference on intelligent robots and systems, vol. 2, 1995
- Whitesides, G.M., Grzybowski, B.: Self-assembly at all scales. *Science* **295**, 2418–2421 (2002)
- Rothermund, P.W.K.: Folding DNA to create nanoscale shapes and patterns. *Nature* **440**, 297–302 (2006)
- Yeh, H.J.J., Smith, J.S.: Fluidic self-assembly for the integration of GaAs light-emitting diodes on Si substrates. *IEEE Photonics Technol. Lett.* **6**, 706–708 (1994)
- Alien Technology. Available: <http://www.alientechology.com> (2011). Accessed 1 Mar 2011
- Saeedi, E., Kim, S., Parviz, B.A.: Self-assembled crystalline semiconductor optoelectronics on glass and plastic. *J. Micromech. Microeng.* **18**, 7 (2008)
- Saeedi, E., Etzkorn, J.R., Parviz, B.A.: Sequential self-assembly of micron-scale components with light. *J. Mater. Res.* **26**, 1 (2011)
- Fukushima, T., Konno, T., Tanaka, T., Koyanagi, M.: Multichip self-assembly technique on flexible polymeric substrate. In: 58th Electronic Components and Technology Conference 2008 (ECTC 2008), pp. 1532–1537 (2008)
- Srinivasan, U., Liepmann, D., Howe, R.T.: Microstructure to substrate self-assembly using capillary forces. *J. Microelectromech. Syst.* **10**, 17–24 (2001)
- Jacobs, H.O., Tao, A.R., Schwartz, A., Gracias, D.H., Whitesides, G.M.: Fabrication of a cylindrical display by patterned assembly. *Science* **296**, 323–325 (2002)
- Knuesel, R.J., Jacobs, H.O.: Self-assembly of microscopic chiplets at a liquid-liquid-solid interface forming a flexible segmented monocrystalline solar cell. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 993–998 (2010)
- Park, K.S., Xiong, X., Baskaran, R., Böhringer, K.F.: Fluidic self-assembly of millimeter scale thin parts on preprogrammed substrate at air-water interface. In: IEEE International Conference on Micro Electro Mechanical Systems (MEMS), pp. 504–507 (2010)
- Fang, J., Böhringer, K.F.: Parallel micro component-to-substrate assembly with controlled poses and high surface coverage. *J. Micromech. Microeng.* **16**, 721–730 (2006)

15. Hoo, J., Baskaran, R., Böhringer, KF.: Programmable batch assembly of microparts with 100 % yield. In: *Transducers*, Denver, pp. 829–832 (2009)
16. Hoo, J., Lingley, A., Baskaran, R., Xiong, X., Böhringer, K.F.: Parallel assembly of 01005 surface mount technology components with 100 % yield. In: *IEEE International Conference on Micro Electro Mechanical Systems (MEMS)*, pp. 532–535 (2010)
17. Ramadan, Q., Uk, Y.S., Vaidyanathan, K.: Large scale microcomponents assembly using an external magnetic array. *Appl. Phys. Lett.* **90**, 172502/1–172502/3 (2007)
18. Fischer, A.C., Roxhed, N., Haraldsson, T., Heinig, N., Stemme, G., Niklaus, F.: Fabrication of high aspect ratio through silicon vias (TSVs) by magnetic assembly of nickel wires. In: *IEEE 24st International Conference on Micro Electro Mechanical Systems 2011(MEMS 2011)*, Cancun (2011)

Self-Assembly of Nanostructures

Wei Lu

Department of Mechanical Engineering,
University of Michigan, Ann Arbor, MI, USA

Synonyms

Self-assembled nanostructures; Self-organized nanostructures

Definition

Self-assembly of nanostructures is a process where atoms, molecules or nanoscale building blocks spontaneously organize into ordered structures or patterns with nanometer features without any human intervention. It is the most promising practical low-cost and high-throughput approach for nanofabrication.

The Concept of Self-Assembly

The term “self-assembly” can be understood from its two components. The first component “self” implies “spontaneous and on its own,” which suggests that it is a process that happens without

human intervention or operation from outside of the system. The second component “assembly” indicates “forming or putting together,” which suggests that the result is a structure built up by lower level building blocks or parts. Self-assembly of nanostructures refers to structures or patterns with nanometer features that form spontaneously from the basic building blocks such as atoms, molecules, or nanoparticles. Cells and living organisms are perfect examples of fairly sophisticated structures self-assembled in nature. These functional assemblies have motivated the study and design of nonliving systems.

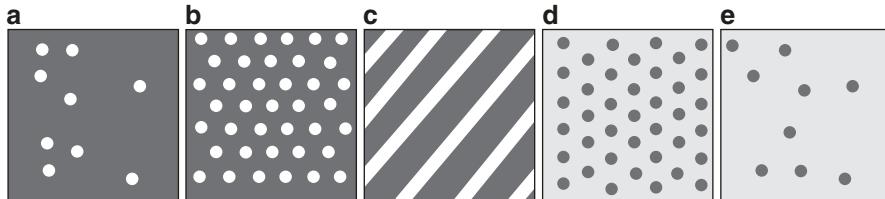
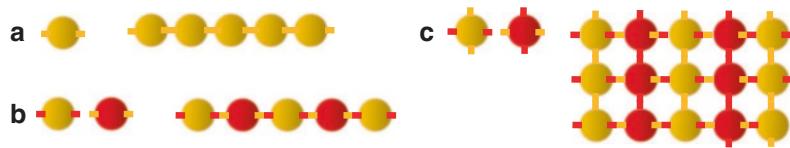
From a scientific point of view, the spontaneous formation of ordered nanostructures from a random disordered state is inherently an intriguing process in any system. The fundamental mechanisms behind these behaviors raise significant research interests. In addition, studying self-assembly in nonliving systems may provide the critical understanding toward decoding the living systems, which is far more complicated. From an engineering point of view, self-assembly as the fundamental principle of a bottom-up approach is possibly the most promising technique for nanofabrication. Thus self-assembly plays a central role in the broad nanotechnology field. The self-assembly of nanostructures may enable a wide range of applications [1], such as nanoelectronic devices, ultrasensitive biosensors, carriers for drug delivery, high capacity lithium batteries, efficient photovoltaic devices, and advanced materials with unique mechanical, electrical, magnetic, or photonic properties.

Discrete System Versus Continuum System

Depending on the type of building blocks, self-assembled systems can be classified into two categories: discrete and continuum. A discrete system uses prefabricated building blocks with fixed sizes and shapes. Figure 1 shows an example, where the building blocks are spherical nanoparticles with engineered bonding sites through surface treatment. When put in a liquid, these nanoparticles collide randomly due to the

Self-Assembly of Nanostructures,

Fig. 1 An example shows the self-assembly of a discrete system



Self-Assembly of Nanostructures, Fig. 2 An example shows the self-assembly of a continuum system. The domain patterns evolve from *dots* to *parallel stripes*, and to *inverted dots* with the increasing coverage of the bright phase

Brownian motion. When two particles are close to each other with matching orientations so that their bonding sites meet, a bond between them forms. A chain structure as shown in Fig. 1a will appear when the system has a single type of nanoparticles. Figure 1b shows a system composed of two types of nanoparticles. A yellow particle has two red bonding sites, indicating that it will bond to two red particles. A red particle has two yellow bonding sites, indicating that it will bond to two yellow particles. This selective bonding leads to a self-assembled chain of alternating particles. Figure 1c shows a case where each particle has four bonding sites. They self-assemble into a network of alternating particle chains. Similar approach can be used to construct three-dimensional structures as well. Various structures can self-assemble by engineering the bonding properties.

In addition to local bonding properties, electric and magnetic fields as well as shear forces and spatial constraints have been used to direct the assembly of nanoparticles and nanorods into different configurations [2]. The electric and magnetic field method induce a dipole-type long-range interaction to act as the driving force to bring nanoparticles together, while the shear force method uses hydrodynamic interactions. The assembly of nanowire arrays is more challenging than that of nanoparticles and nanorods due to their highly anisotropic shape. Nanowire

self-assembly usually results in partially ordered, small superlattices. This issue has been addressed by several new methods to direct the process, including the use of microfluidic channels and electric fields [3, 4].

In contrast, a continuum system exploits the spontaneous formation of nanoscale domains. Examples include self-assembled domain patterns in binary monolayers [5], block copolymers [6], and organic molecular adsorbates on metal surfaces [7, 8]. A binary monolayer on an elastic substrate may separate into two phases, and self-assemble into ordered patterns, such as triangular lattice of dots, parallel stripes, or serpentine stripes. The feature size is on the order of 1–100 nm, and stable against annealing. Block copolymers are polymers consist of at least two chemically distinct, immiscible polymer fragments that are joined by a covalent bond. These systems have been shown to develop a variety of regular domain patterns via phase separation [9, 10]. The size and the period of the structures are typically on the order of 10–100 nm, depending on the conditions of preparation and the relative chain lengths of the participating polymers.

Figure 2 shows an example of domain patterns formed by two phases. A notable feature is the dependence of the pattern on the average concentration. Take Cu and Pb monolayer on Cu (111) substrate as an example. The Pb and Cu mixture monolayer on a Cu substrate forms

two phases: a Pb overlayer and a disordered Pb–Cu surface solution. The two phases in the monolayer self-assemble into ordered, nanoscale patterns. When the average concentration of Pb atoms increases, the system can experience a series of patterns from (a) to (e), where the bright phase is Pb while the dark phase is the Pb–Cu surface alloy. When the average concentration is very low, isolated dots are obtained, as shown in Fig. 2a. The bright phase is embedded in the continuous matrix of the dark phase. No long-range ordering is observed. When the average concentration increases, the area covered by the bright dots increases. The dots order and form patterns close to a triangular lattice, as shown in Fig. 2b. When the bright phase occupies roughly half of the surface, stripe structure as shown in Fig. 2c forms. “Inverted” dots are observed when the bright phase dominates, as shown in Fig. 2e.

A continuum system offers several unique features. For instance, domains and their patterns self-assemble simultaneously, so that there is no need to pre-synthesize the building blocks; a significant degree of process flexibility and control can be achieved; and the approach may be applied to diverse systems.

Static Self-Assembly Versus Dynamic Self-Assembly

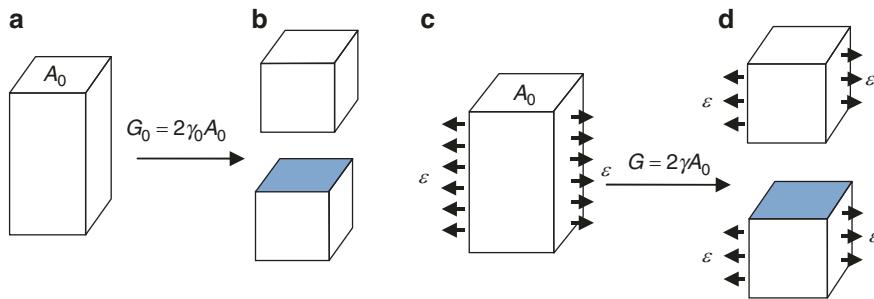
Depending on the nature of interactions, self-assembly systems can be classified into two categories: static and dynamic. Static self-assembly is a process driven by energy minimization to form static structures which are at global or local equilibrium. Many examples fall into this category, such as colloidal crystals, microphase separation, or spontaneous domain patterns. Dynamic self-assembly involves interactions that only occur when the system is dissipating energy. The structure is determined by an energetic minimum relying on an influx of energy to the system. When the energy flux stops, the minimum configuration does not exist anymore, leading to disintegration of the structure. For instance, rotating nanodisks in a liquid can generate hydrodynamic interactions due to the flow of liquid. When the disks

are close to each other, the hydrodynamic interaction causes them to form a nicely ordered hexagonal pattern with a regular inter-disk spacing. When the rotation stops, the interaction disappears and the pattern vanishes. Another example is the oscillating reaction–diffusion reaction, where an evolving pattern develops from a quiescent medium upon the influence of stimuli. A living organism is possibly the most typical example of dynamic self-assembly.

Compared to static self-assembly, dynamic self-assembly is even less understood [11]. This process involves interacting building blocks that can adapt or react to a chemical or physical stimulus from the environment. Unlike the forces that drive static self-assembly through a reduction of the free energy toward equilibrium, the interactions in dynamic self-assembly may drive the formation of structures and patterns away from the thermodynamic equilibrium sustained by a continuous energy input. Such patterns can thus be sensitive to external stimuli and adaptive in response to the surrounding conditions [12].

Forces That Drive Self-Assembly

Forces of different physical origins may contribute to the interactions between the building blocks. The short-range interaction is generally controlled by forces such as ionic, covalent, or metallic bonding; hydrogen bonding; and van der Waals forces. The long-range interaction has an effective range much longer than the atomic distance, and can originate from elasticity, electrostatic or magnetic field, colloidal and capillary forces, or hydrodynamic interaction. The short-range interaction alone typically leads to either a homogenous structure as shown in Fig. 1a, or a structure with only local modulation whose wavelength is on the order of the size of the building blocks, such as those in Fig. 1b, c. The long-range interaction is essential to form characteristic feature sizes larger than the dimension of the building blocks. Thus long-range interaction is especially important for the self-assembly in a continuum system to form nanoscale feature size, which is much larger than the building block of a single atom or molecule.



Self-Assembly of Nanostructures, Fig. 3 Schematic representation illustrates the concept of surface energy and surface stress

In a discrete system of nanoscale building blocks, the feature size is determined by the competition between the attractive and repulsive forces. Thus force balance analysis is a useful approach. While such a concept may still be applied to a continuum system, the connection between the domain size and their interaction is much less direct. Therefore the energy method is usually used to analyze a continuum system. The feature size is determined by two competing actions: a coarsening action which tends to increase the domain size and a refining action which tends to reduce the domain size. The following discusses a few representative forces and their roles in self-assembly.

Surface Stress

Surface stress plays crucial roles in a variety of surface phenomena in solids. To explain the idea, consider a uniform and infinitely large solid shown in Fig. 3a. The body is unstrained and is taken as the reference state. In Fig. 3b, the body is cut into two parts, and the atoms on the surfaces are allowed to find their equilibrium configuration. The energy in state (b) is higher than that in state (a). The energy difference between the two states, G_0 , can be written as $G_0 = 2\gamma_0 A_0$, where γ_0 is the excess free energy per unit area owing to the existence of the surface, and A_0 is the surface area. G_0 is called surface energy. γ_0 is usually called surface energy density, which is the reversible work per unit area to create a surface.

Figure 3c is the same solid in (a) but under strain ε . In Fig. 3d, the strained body is cut into two parts. The surface energy can be written as $G = 2\gamma A_0$. The two energies, G and G_0 , are different. The former depends on the strain, i.e., $G = G(\varepsilon)$. Accordingly, the surface energy density also depends on the strain state, $\gamma = \gamma(\varepsilon)$. Hence the surface energy depends on both the created surface area and the strain state, namely, $d(\gamma A_0) = A_0 f d\varepsilon$, where $f = \partial\gamma/\partial\varepsilon$ is called surface stress. Surface stress becomes a tensor when generalized to biaxial strains.

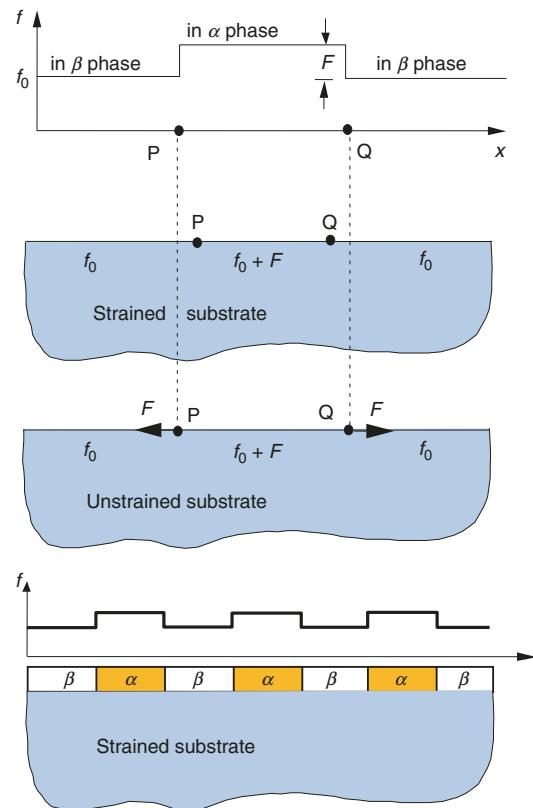
The physical origin of surface stress is related to the difference of bonding between atoms at the surface and the interior atoms. A typical approach to measure surface stress is the cantilever bending method. Upon deposition of atoms on the top surface, surface stress causes the cantilever to bend. By measuring the curvature and some geometric parameters, such as the thickness of the cantilever, one can calculate surface stress. The surface stress of a domain depends on its composition. For a binary epilayer, a composition modulation will cause surface stress nonuniformity.

Consider a thin binary layer that grows epitaxially on a solid substrate with fixed thickness, say a monolayer. The two species can relocate by diffusion within the layer. Several physical ingredients contribute to the spontaneous domain formation behavior. Phase separation is a commonly observed phenomenon in various material systems, which can be explained in terms of the free energy of mixing. Consider a mixture of two atomic species A and B, and define concentration

by the fraction of one species. The free energy density is given by a double well function of the concentration. A tangent line contacts the function at two concentrations, corresponding to the equilibrium A-rich and B-rich phases. A mixture with an average concentration between these two concentrations will separate into the two phases to reduce the energy. The atoms at the phase boundaries have excess free energy. To reduce the free energy, the total area of the phase boundaries must reduce. Consequently, atoms leave small domains, diffuse in the matrix, and join large domains. Over time the small domains disappear, and the large ones become larger. Thus phase boundary energy causes phase coarsening, which cannot form stable patterns. The observed stable periodic patterns suggest that there must be a refining action in the regular nanoscale structures. This action is provided by the surface stress.

Figure 4 shows an important concept: substrate deformation allows the nonuniform surface stress to reduce the total energy. The surface stress is f_0 everywhere except for the portion of the surface between points P and Q, where the surface stress steps up by F . The nonuniformity causes P and Q to move toward each other, giving rise to a strain field in the substrate. Imagine that a pair of forces are applied to move the points P and Q away from each other. When the applied forces reach the magnitude F , the substrate becomes unstrained. Because the forces do work to the body, the unstrained state has a higher energy than the strained state. The energy change depends only on F , but not on f_0 . Whether the overall surface stress is tensile, compressive, or vanishing does not make any difference to the energy reduction. When the domain size is refined, more pairs of F will contribute so that the energy is reduced further. Consequently, the elastic energy in the film-substrate composite tends to refine the domains. The refinement, however, adds more domain boundary, which increases the phase boundary energy. The phase boundary energy tends to coarsen the domains. The two competing actions – refining and coarsening – can select an equilibrium domain size.

These physical ingredients have been incorporated into an energy framework to simulate

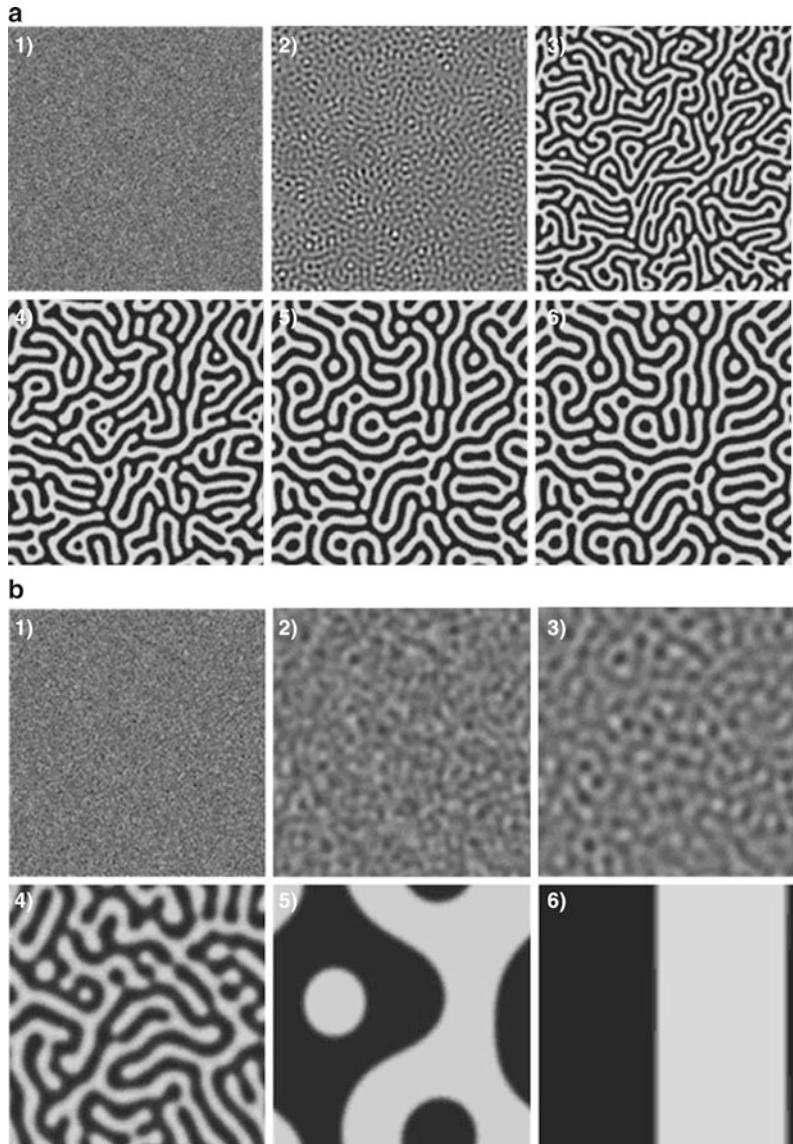


Self-Assembly of Nanostructures, Fig. 4 Substrate deformation allows the nonuniform surface stress to reduce the energy, and surface stress provides a refining action

various nanostructures [13]. Figure 5 clearly shows the refining effect of the surface stress. The average concentration is 0.5. Each phase takes half of the surface area. In sequence (a), shortly after phase separation, the two phases form serpentine stripes. The width of the stripes stabilizes very fast. From $t = 1,000$ to $t = 10^6$, the widths are almost invariant. In contrast, in sequence (b), the two phases try to increase their sizes as much as possible. The system finally evolves into a state that one phase takes half of the calculation cell and the other phase takes the rest half. This reproduces the classical spinodal decomposition. The pattern depends on the average concentration. Changing the average concentration away from 0.5 can generate dots instead of stripes. Anisotropic surface stress induces interesting patterns such as parallel stripes and herringbone structures [14].

Self-Assembly of Nanostructures,

Fig. 5 Simulation starts with a random initial condition with an average concentration of 0.5. (a) The phases form nanoscale serpentine stripes with surface stress. (b) Spinodal decomposition without surface stress. The phases always coarsen. The times are (1) $t = 0$, (2) $t = 10$, (3) $t = 100$, (4) $t = 1,000$, (5) $t = 10^5$, (6) $t = 10^6$ (Adapted from Ref. [14])



Guided or Tempered Self-Assembly

Guided or templated self-assembly uses a coarse scale external field to influence the fine scale self-assembly behavior. An essence of the approach is that the wavelength of the guiding field or template can be much larger than the nanoscale structures to be formed. Consequently, preparing the coarse scale field involves low cost and can be easily applied to a very large area. The guiding field can be electrostatic, geometric pre-patterns, surface chemistry, or elastic field [15].

Figure 6 shows an example to guide the surface patterns with strain field on the substrate surface. Surface strain is applied in the black region in Fig. 6a. Three cases are shown: (1) no guidance, (2) guiding strain in wide stripes, and (3) guiding strain in wavy stripes. In practice, there are many ways to induce an elastic field on the substrate surface. In addition to direct mechanical loading, pre-patterning a substrate with different materials by photolithography or applying an electric field to a substrate embedded with piezoelectric particles produce diverse well-defined strain fields.

Self-Assembly of Nanostructures,

Fig. 6 Guided self-assembly with strain field on the substrate surface. (a) A constant strain is applied in the *black region* to guide the self-assembly on the substrate surface. (b) Patterns for an average concentration of 0.3. (c) Patterns for an average concentration of 0.5 (Adapted from Ref. [15])

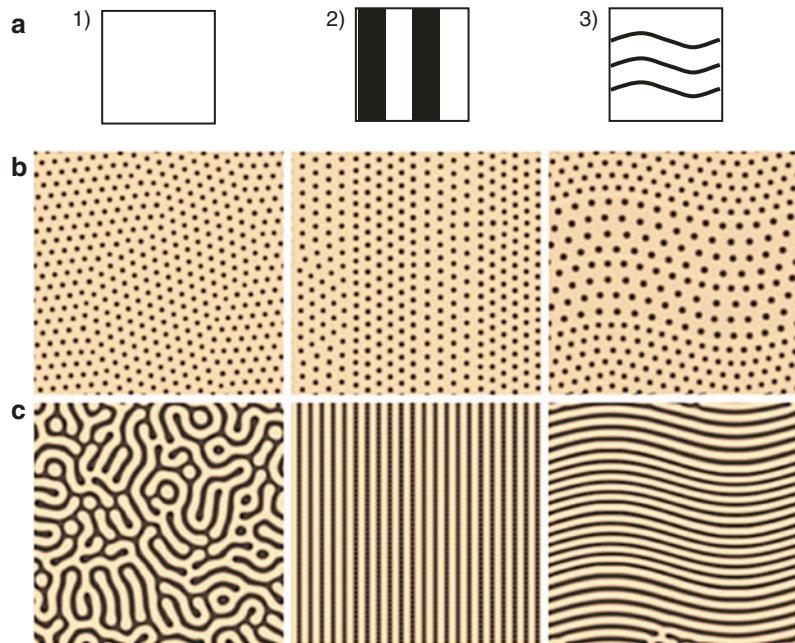


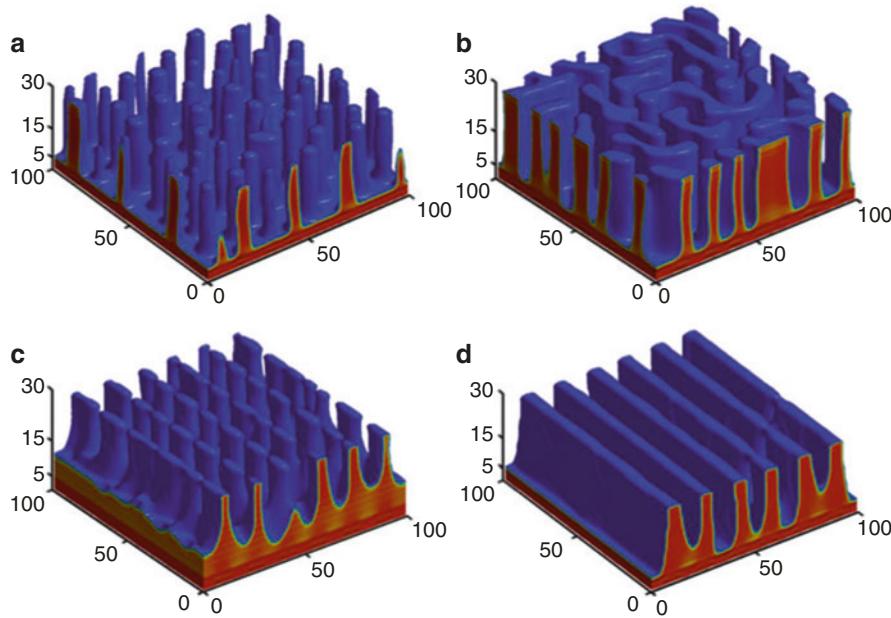
Figure 6b shows patterns for an average concentration of 0.3. The monolayer separates into two phases and evolves into a multi-domained triangular lattice of dots when there is no guidance. The middle column shows that in case [2], the monolayer evolves into pattern colonies. The dots form an almost perfect triangular lattice within each colony, and line up with the close packed direction parallel to the edge of the stripes. The lattices in the white and black stripe areas are self-contained, maintaining their own lattice spacings. The right column shows that in case [3], the dots orientate themselves along the wavy edges of the three curves. The narrow curves preclude the formation of any dots inside. Similar guiding effect is also shown for an average concentration of 0.5 in Fig. 6c. These results suggest that external loading can effectively change size, shape, and orientation of self-assembled surface structures. Similar guided self-assembly has also been achieved by surface chemistry [16].

Figure 7 shows another example where electric field is used to guide self-assembly. A three-dimensional model has been developed to allow the simulation of the entire self-assembly process and electric field design [17]. It is shown that a melted thin polymer film subjected to an

electrostatic field may lose stability at the polymer-air interface, leading to uniform self-organized pillars emerging out of the film surface. The film thickness can affect the patterns formed. With patterned electrodes, parallel stripes replicating an electrode pattern emerge.

Electric Dipoles

An adsorbate molecule usually carries an electric dipole. Even if the molecules are nonpolar, the act of binding onto a substrate breaks the symmetry and causes the formation of dipoles. A molecule can be engineered to carry a large electric dipole moment by incorporating a polar group. Electric dipole interaction can give rise to domain patterns. Examples include Langmuir films at the air–water interface, ferrofluids in magnetic fields, and organic molecules on metal surface. Despite the difference of these systems, similar phenomenology and mechanism can be identified. The adsorbed molecules are mobile. Domains coarsen to reduce the domain boundary and refine to reduce the dipole interaction energy. The competition leads to equilibrium patterns. This mechanism may be used to make two-dimensional



Self-Assembly of Nanostructures, Fig. 7 Self-assembled morphological patterns of a melted thin polymer film induced by two electrodes parallel to the initial flat film surface, one above the film and the other beneath the film. (a) Induced nanopillars in a thin film. (b) A relatively

thick film. The film continues to evolve after touching the ceiling electrode and then spread. (c) Patterned top electrode causes the nanopillars to line up. (d) Replication of the top electrode pattern (Adapted from Ref. [17])

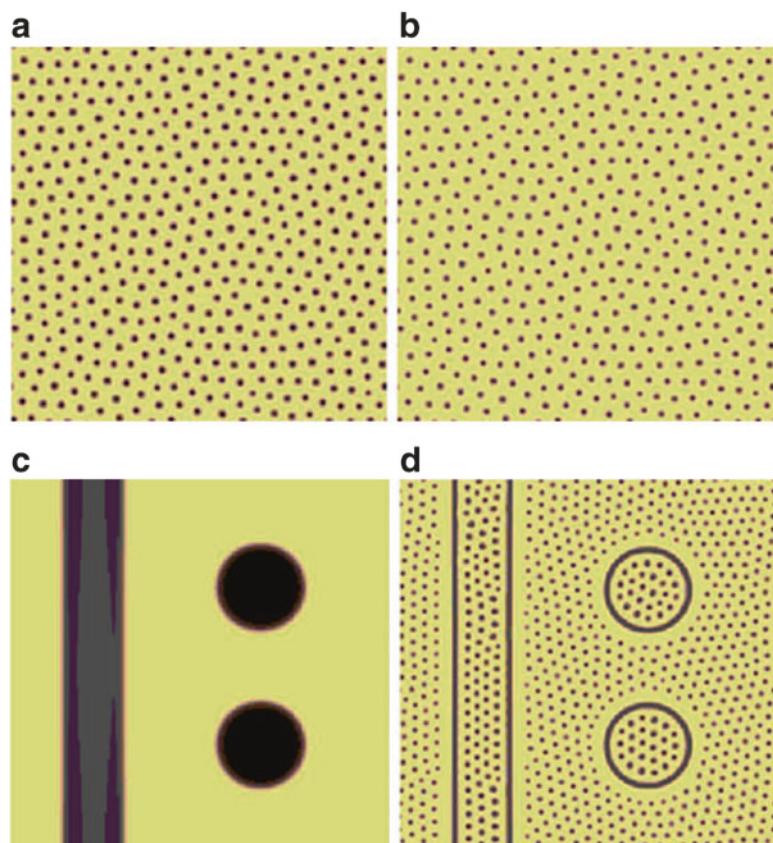
nanostructures. Specifically, molecule monolayers composed of electric dipoles can be manipulated with an electric field induced by an AFM tip, a ceiling above the layer, or an electrode array in the substrate.

Multilayers of dipole molecules have the capability for making complex structures, especially the formation of nanointerfaces and three-dimensional nanocomposites. Studies have revealed self-alignment between layers, reduction of feature sizes, and guided self-assembly by layer-layer interaction and embedded electrodes [18]. Figure 8 gives an example. Each layer is composed of two kinds of molecules with different dipole moments, which form two domains. Figure 8a shows that the first layer self-assembles into a triangular lattice of dots at an average concentration of 0.3. The dots have uniform size and form multiple grains. Figure 8b shows the second layer pattern at an average concentration of 0.2 and grows on top of the pattern in Fig. 8a. The second layer evolves from a completely different random initial condition. The dots of the second

layer stay at exactly the same positions as those in Fig. 8b, suggesting the anchoring effect of the first layer. The dot size of the second layer is smaller due to lower average concentration. The observations suggest that the first layer determines the ordering and lattice spacing, while the second layer determines the feature size. A scaling down of size can be achieved via multilayers. The interesting behavior suggests a potential fabrication method. In addition to self-assembly, the first layer pattern can be defined by embedded electrodes, proximal probe technique, or nanoimprinting. Figure 8c shows the first layer pattern with the application of a high voltage, which sweeps off any self-assembled features so that the monolayer replicates the voltage pattern. The second layer evolves into a pattern shown in Fig. 8d. The dots align themselves along the edges of the first layer pattern, and form triangular lattice. It is interesting to note the formation of pairing dark and white lines following the contour of the first layer pattern. Two nearby regions separated by the lines have different preference for

**Self-Assembly
of Nanostructures,**

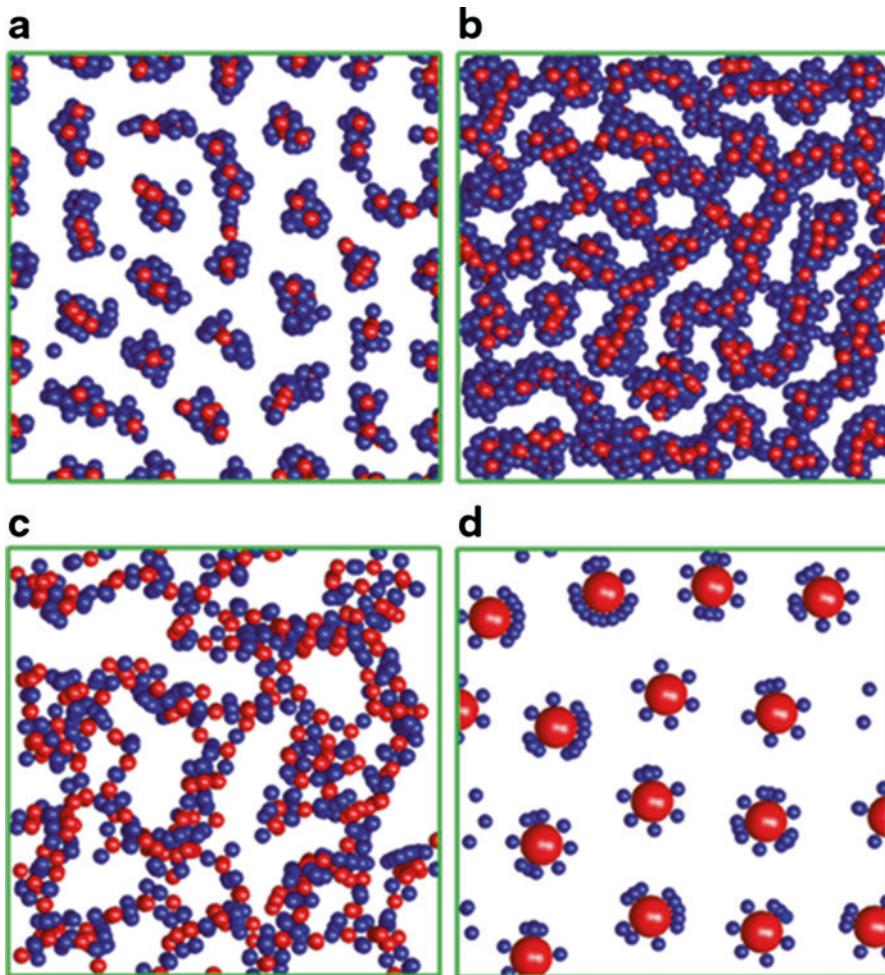
Fig. 8 Self-assembled patterns by electric dipole interaction. (a) First layer, average concentration 0.3. (b) Second layer on top of (a). (c) First layer pattern replicating the electrode pattern. (d) Second layer on top of (c) (Adapted from Ref. [18])



the two dipole species. This causes local accumulation and depletion of the two dipoles, resulting in the observed phenomena.

Another example of electric dipole interaction is the self-assembly of nanoparticles [19]. Under an applied electric field, each particle will acquire an effective induced dipole. The dipole interaction causes particles to move in a liquid medium and line up into chains following the electric field direction. The formation of an ordered three-dimensional structure is further affected by the interaction between the chains. The self-assembly of binary nanoparticles can lead to a wide class of nanocomposite materials with properties not attainable by a single particle component. Figure 9a shows a top view of the structure. Behind each visible particle shown in the figure is a particle chain lining up. The two kinds of particles are denoted by red and blue colors. They have the same diameter and are both more polarizable than the medium, while the red

particle is more polarizable than the blue one. The volume fraction of the particles is 10.2 %. Under these situations the particles assemble into isolated columns with a core-shell configuration. The core is composed of the more polarizable red particles, while the shell is composed of the blue ones. From an energetic point of view, the attraction between two red particle chains is stronger than the interaction between a red particle chain and a blue particle chain. The latter is strong than the attraction between two blue particle chains. As a result, the red chains aggregate to form columns. The blue chains tend to get close to the red columns as much as possible, leading to the formation of shells. It is exciting to note that this self-organized functionally gradient structure offers a gradual transition of the permittivity from that of the core to that of the medium. This approach could be potentially very useful to construct functional gradient nanocomposites from bottom up.



Self-Assembly of Nanostructures, Fig. 9 *Top* view of structures self-assembled in a system of binary nanoparticles. (a) Gradient core-shell structure. (b)

A continuous structure with isolated holes in between. (c) Network structure. (d) A single layer (Adapted from Ref. [19])

Figure 9b shows a higher volume fraction of particles, 30.6 %. The system evolves into a continuous structure with isolated holes in between. The blue particles enclose the red particles, and form the peripheral regions around the holes. The particle columns demonstrate local BCT structures. Another common feature is that same kind particles tend to aggregate. Figure 9c demonstrates the structure when one kind of particle is more polarizable than the medium while the other is less polarizable. The particles also form pure chains along the field direction with each chain composed of single kind particles. A distinct

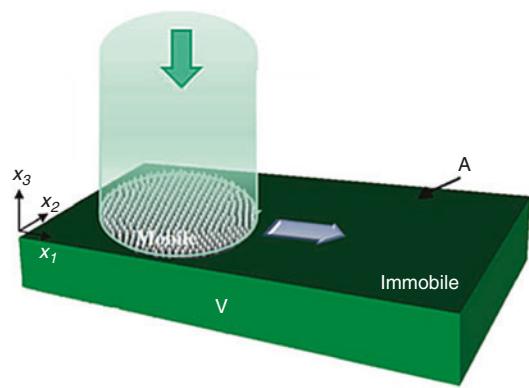
feature in Fig. 9c is that the red and blue chains are highly dispersed and form a network. This morphology is in contrast to that in Fig. 9a, where same color chains aggregate.

Assembling a single layer of nanoparticles on a substrate has many potential applications. A relevant simulation is shown in Fig. 9d for a layer of two kinds of particles. The repulsion between the red particles causes them to form a nicely ordered triangular lattice. The attraction between the blue and red particles causes the formation of blue rings surrounding the red cores.

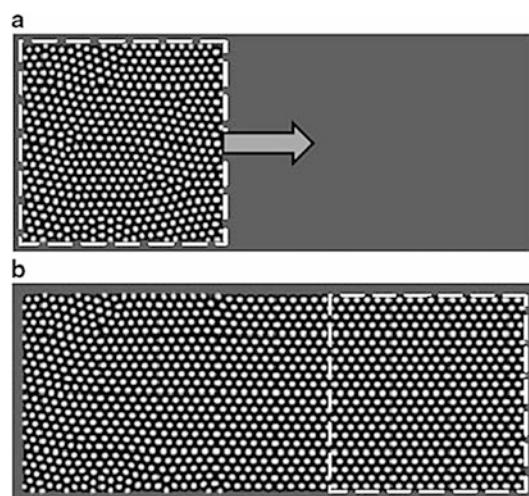
Sequential Activation of Self-Assembly

The lack of long-range order has been a major challenge for self-assembled nanostructures since high regularity is crucial to many applications. A general kinetics-based template-free approach has been developed recently to grow nanostructured superlattices over a very large area [20]. The essence is sequential activation of self-assembly. The idea is illustrated in Fig. 10. A binary monolayer is used as an example. Typically multiple grains of dots will form due to simultaneous self-assembly at different locations, producing a pattern lack of long-range order. Here self-assembly is first activated in a finite mobile region, where atoms are allowed to diffuse and form domain patterns. This initial mobile region will serve as a “seed.” The seed does not need to have a perfect lattice. Then the mobile region is shifted like scanning. The self-assembly in the newly activated region will be influenced by the pattern already formed in the seed. In experiments this process can be achieved by laser or ion beam scanning to control the local temperature, so that diffusion is activated sequentially at each spot along the scanning path. This sequential activation will lead to a large long-range ordered domain pattern even if started with an imperfect seed. The pattern quickly improves and converges to a perfect superlattice along with the sequential activation.

Figure 11 demonstrates the growth of long-range ordered superlattices from a seed. Rather than using a perfect lattice directly, the seed is grown on site. Take Fig. 11a as an example. The constraint of kinetics leads to a square seed composed of dots. The seed is larger than the size of a typical single grain. Therefore defects such as misalignment and multiple grains appeared in this seed. Next the mobile region is shifted to the right. This process is called scanning. The shift distance in each step is much smaller compared to the seed size, which creates a continuous scanning effect. The scanning velocity is chosen to ensure ample time for new domains to develop. Figure 11b shows the structure after scanning over the width of the calculation cell, which forms a band of nicely ordered hexagonal

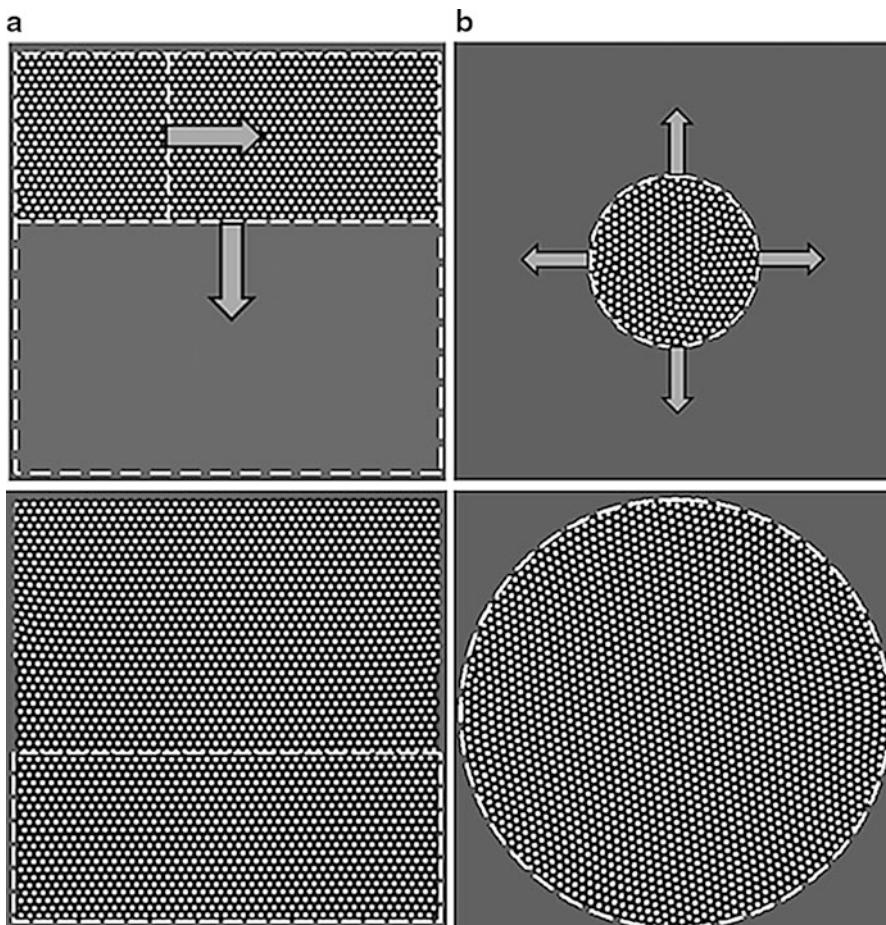


Self-Assembly of Nanostructures, Fig. 10 A schematic of sequential activation of self-assembly. Self-assembly is first activated in a finite mobile region, where atoms are allowed to diffuse and form domain patterns. This initial mobile region serves as a “seed.” The mobile region is then shifted like scanning. Self-assembly in the newly activated region will be under influence of patterns already formed in the seed (Adapted from Ref. [20])



Self-Assembly of Nanostructures, Fig. 11 Growth of superlattice from seeds. (a) A square seed. (b) A band of nicely ordered hexagonal superlattice formed after scanning over the width. The lattice improved to perfect along with the scanning, demonstrating tolerance of defects in the seeds (Adapted from Ref. [20])

superlattice. Noticeably, the lattice improves to perfection along with the scanning, demonstrating the tolerance of defects in the seeds. High output is essential to nanostructure applications.



Self-Assembly of Nanostructures, Fig. 12 Two schemes for scaling-up growth. (a) Alternate the scanning directions, using the superlattice created in previous step as

a large seed. (b) Increase the size of the mobile in two-dimension (Adapted from Ref. [20])

Two schemes have been proposed to facilitate large-scale fabrication. (1) As shown in Fig. 12a, the scanning directions are alternated between left-right scanning and up-down scanning, using the superlattice created in each previous step as a large seed. This scheme allows the growth rate (area of lattice created per unit time) to increase exponentially with time, greatly accelerating large area fabrication. (2) As shown in Fig. 12b, the size of the mobile area is increased in two dimensions, rather than scanning along one direction. This scheme allows the growth rate to be quadratic of the scanning velocity.

Cross-References

- Charge Transport in Self-Assembled Monolayers
- Nanomaterials for Electrical Energy Storage Devices
- Nanomaterials for Excitonic Solar Cells
- Nanoparticles
- Nanostructured Functionalized Surfaces
- Nanotechnology
- Self-Assembled Monolayers
- Self-Assembly
- Self-Assembly for Heterogeneous Integration of Microsystems

References

1. Lu, W., Sastry, A.M.: Self-assembly for semiconductor industry. *IEEE Trans. Semicond. Manuf.* **20**, 421–431 (2007)
2. Pileni, M.P.: Nanocrystal self-assemblies: fabrication and collective properties. *J. Phys. Chem. B* **105**, 3358–3371 (2001)
3. Huang, Y., Duan, X., Wei, Q., Lieber, C.M.: Directed assembly of one-dimensional nanostructures into functional networks. *Science* **26**, 630–633 (2001)
4. Smith, P.A., Nordquist, C.D., Jackson, T.N., Mayer, T.S., Martin, B.R., Mbindyo, J., Mallouk, T.E.: Electric-field assisted assembly and alignment of metallic nanowires. *Appl. Phys. Lett.* **77**, 1399–1401 (2000)
5. Plass, R., Last, J.A., Bartelt, N.C., Kellogg, G.L.: Nanostructures – self-assembled domain patterns. *Nature* **412**, 875 (2001)
6. Fasolka, M.J., Mayes, A.M.: Block copolymer thin films: physics and applications. *Annu. Rev. Mater. Res.* **31**, 323–355 (2001)
7. Love, J.C., Estroff, L.A., Kriebel, J.K., Nuzzo, R.G., Whitesides, G.M.: Self-assembled monolayers of thiolates on metals as a form of nanotechnology. *Chem. Rev.* **105**, 1103–1169 (2005)
8. Schneider, K.S., Lu, W., Owens, T.M., Fosnacht, D.R., Banaszak Holl, M.M., Orr, B.G.: Monolayer pattern evolution via substrate strain-mediated spinodal decomposition. *Phys. Rev. Lett.* **93**, 166104 (2004)
9. Krausch, G., Magerle, R.: Nanostructured thin films via self-assembly of block copolymers. *Adv. Mater.* **14**, 1579–1583 (2002)
10. Lopes, W.A., Jaeger, H.M.: Hierarchical self-assembly of metal nanostructures on diblock copolymer scaffolds. *Nature* **414**, 735–738 (2001)
11. Whitesides, G.M., Boncheva, M.: Beyond molecules: self-assembly of mesoscopic and macroscopic components. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 4769–4774 (2002)
12. Ozin, G.A., Hou, K., Lotsch, B.V., Cademartiri, L., Puzzo, D.P., Scotognella, F., Ghadimi, A., Thomson, J.: Nanofabrication by self-assembly. *Mater Today* **12**, 12–23 (2009)
13. Lu, W., Suo, Z.: Dynamics of nanoscale pattern formation of an epitaxial monolayer. *J. Mech. Phys. Solids* **49**, 1937–1950 (2001)
14. Lu, W., Suo, Z.: Symmetry breaking in self-assembled monolayers on solid surfaces: anisotropic surface stress. *Phys. Rev. B* **65**, 85401 (2002)
15. Lu, W., Kim, D.: Engineering nanoparticle self-assembly with elastic field. *Acta Mater.* **53**, 3689–3694 (2005)
16. Lu, W., Kim, D.: Patterning nanoscale structures by surface chemistry. *Nano Lett.* **4**, 313–316 (2004)
17. Kim, D., Lu, W.: Three-dimensional model of electrostatically induced pattern formation in thin polymer films. *Phys. Rev. B* **73**, 035206 (2006)
18. Lu, W., Salac, D.: Patterning multilayers of molecules via self-organization. *Phys. Rev. Lett.* **94**, 146103 (2005)
19. Park, J., Lu, W.: Self-assembly of functionally graded nanoparticle structures. *Appl. Phys. Lett.* **93**, 243109 (2008)
20. Zhao, Z., Lu, W.: Growing large nanostructured superlattices from a continuum medium by sequential activation of self-assembly. *Phys. Rev. E* **83**, 041610 (2011)

Self-Cleaning

► Lotus Effect

Self-Energy and Excitonic Calculations in Many-Body Systems

► Electronic and Optical Properties of Oxides Nanostructures by First-Principles Approaches

Self-Healing Materials

► Self-Repairing Materials

Self-Organized Layers

► Self-Assembled Monolayers for Nanotribology

Self-Organized Nanostructures

► Self-Assembly of Nanostructures

Self-Regeneration

► Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components

Self-Repairing Materials

Michael Nosonovsky

Department of Mechanical Engineering,
University of Wisconsin-Milwaukee, Milwaukee,
WI, USA

Synonyms

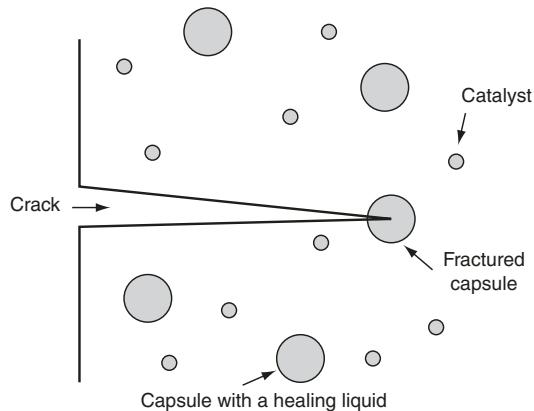
Self-healing materials

Definition

Self-repair or self-healing is the ability of a material (usually, a composite or nanocomposite) to partially or completely repair damage, such as voids and cracks, occurring during its service lifetime [1, 2]. Self-repair is achieved by embedding a nano-, micro-, or multi-scale repair mechanism into the structure of the material (Fig. 1). Autonomous self-healing refers to the ability to repair damage without any external activation. Nonautonomous self-repair refers to healing mechanisms, which require external activation, such as heating the material.

Occurrence and Key Findings

Self-healing is related to the general ability of many systems for self-organization. To facilitate self-organization, usually a special nano-, micro-, or multi-scale structure of a material is needed. Self-healing in biological objects is a source of inspiration for this new research area. Most living tissues or organisms can regenerate and heal themselves, provided the incurred damage is small or moderate. Most engineered materials, however, deteriorate with time irreversibly due to wear, brittle fracture, fatigue, creep, and other modes of degradation, which limits the life of various components and sometimes causes catastrophic damage. Biological mechanisms of healing are very complex and cannot be directly



Self-Repairing Materials, Fig. 1 Self-repair by embedding microcapsules with healing agent

borrowed for artificial materials. Therefore, it is preferable to speak about bio-inspired, rather than biomimetic, self-healing materials.

Polymers

Self-healing has been applied most successfully in polymers since they have relatively high rates of diffusion and plasticity due the presence of the cross-molecular bonds. One way to create self-healing polymers is to use thermosetting polymers that have the ability to cure (toughening or hardening by cross-linking of polymer chains), such as the thermosetting epoxy. Epoxy is a polymer formed by a reaction of an epoxide resin with polyamine hardener. Epoxy can serve as a healing agent that is stored within thin-walled inert brittle macrocapsules embedded into the matrix along with a catalyst or hardener. The catalyst or hardener is also embedded in the matrix, but separately from the healing agent. When a crack propagates, the capsules fracture, the healing agent is released and propagates into the crack due to capillarity. Then the healing agent mixes with the catalyst embedded in the matrix, which triggers the cross-linking reaction and hardening of the epoxy that seals the crack [3].

A different approach involves thermoplastic polymers with various ways of incorporating the healing agent into the material. In this approach, heating is often required to initiate healing [3]. A self-repair method involving two concentric cylinders of conducting material filled with a liquid

solution containing electromagnetic particles of polystyrene or silica was suggested. When damage occurs, voltage is applied between the inner pipe and outer pipe, the current density naturally increases at the location of the damage, this increase in current density causes particle coagulation at the damage site, which in a way to heal the damage.

Ceramics

Besides the polymers, ceramic self-healing materials are being developed. For example, a concrete composite was produced with glass fibers containing an air-curing sealant embedded in the concrete matrix. This composite exhibited the self-healing behavior, but it suffered from a significant (10–40 %) loss of stiffness compared with standard concrete due to fibers. This is a typical situation, when a compromise between self-healing and mechanical properties should be sought. In another project involving self-healing ceramics, researchers have studied the crack-healing behavior and mechanical properties of a mullite composite toughened by the inclusion of 15 % (by volume) SiC whiskers. Self-healing ceramic materials often use oxidative reactions because the volume of oxide exceeds the volume of the original material, and therefore, products of these reactions can be used to fill small cracks [3]. Self-healing nanocomposites constitute another area of research.

Metallic Materials

It is much more difficult to heal metallic materials, than polymers, because metallic atoms are strongly bonded and have small volumes and low diffusion rates. Currently, there are three main directions which have been taken in the development of self-healing metallic systems. First is the formation of precipitates at the defect sites that immobilize further growth until failure. This mechanism is called sometimes “damage prevention” because it prevents the formation of voids by the diffusion of the precipitate. The atomic transport of matter to voids and defects is provided by a supersaturated solid solution in alloys having decreasing solid solubility of solute elements with decreasing temperature (e.g.,

Al–Cu). Such an “under-aged” alloy, when quenched from high temperature, becomes supersaturated or metastable. However, in order to facilitate precipitation of the solute, heterogeneous nucleation sites are needed. Sites with high surface energy, such as voids, defects, grain boundaries, and free surfaces, become nucleation sites. The driving mechanism for the diffusion is the excess surface energy of microscopic voids and cracks that serve as nucleation centers of the precipitate that plays the role of the healing agent. As a result, the newly formed void is sealed before it grows and thus minimizing the creep and fatigue [1, 2].

Second is reinforcement of an alloy matrix with microfibers or wires made of a shape-memory alloy (SMA), such as Nitinol (NiTi). SMA wires have the ability to recover their original shape after some deformation has occurred if they are heated above certain critical temperature. If the composite undergoes crack formation, heating the material will activate the shape recovery feature of the SMA wires which then shrink and close the cracks [1, 2].

The third approach is to use a healing agent (such as an alloy with a low-melting temperature) embedded into a metallic solder matrix, similarly to the way it is done with the polymers. However, encapsulation of a healing agent into a metallic material is much more difficult task than in the case of polymers. The healing agent should be encapsulated in microcapsules which serve as diffusion barriers and which fracture when crack propagates [1, 2].

Self-healing mechanisms in metals are summarized in Table 1, showing typical materials for the matrix and reinforcement, parameters that characterize microstructure, degradation, and healing, characteristic length scales for the degradation and healing mechanisms, the type of phase transition involved into the healing and what property is improved in the self-healing alloy, the nature of the healing force and details of healing mechanisms will be discussed in the consequent section.

Surface Healing

Surface is often the most vulnerable area of a material. Several approaches have been

Self-Repairing Materials, Table 1 Self-healing mechanisms in metals

Mechanism	Precipitation	SMA reinforcement	Healing agent encapsulation
Type	Damage prevention (solid)	Damage management (solid, possibly liquid-assisted)	Damage management (liquid-assisted)
Matrix material	Al–Cu, Fe–B–Ce, Fe–B–N, etc.	Sn–Bi, Mg–Zn	Al
Reinforcement materials	—	NiTi	Sn–Pb
Microstructure parameter (ψ)	Solute fraction	Concentration of microwires	Concentration of microcapsules or low-melting-point alloy
Degradation measure (ξ)	Volume of voids	Volume of voids	Volume of voids
Healing measure (ζ)	Amount of precipitated solute	SMA strain	Amount of released healing agent
Characteristic length of degradation	Void size (microscale)	Void/crack size (macroscale)	Void/crack size (macroscale)
Characteristic length of the healing mechanism	Atomic scale (atomic diffusion)	Microwires diameter (macro or microscale)	Microcapsule size (microscale)
Phase transition involved	Solute precipitation	Martensite/austenite	Solidification of the solder
Healing temperature	Ambient	Martensite/austenite transition	Melting of the low-melting-point alloy
Property improved	Creep resistance	Restored strength	Restored strength and fracture toughness

suggested to for fixing surface damage. One approach is to mimic the healing of skin that has a network of vessels and veins so that cutting the skin triggers the blood flow, its coagulation, and sealing of the cut. The vascular network is a tree-like hierarchical structure that provides a uniform and continuous distribution of fluids throughout the material volume. The vascularization has been successfully used for polymeric composites [4–6].

A biomimetic approach combining self-healing with the Lotus-effect has been suggested, mimicking healing of a leaf surface. In that approach, micro-reservoirs with coating liquid (e.g., wax paraffin) were attached to the surface with the supply of coating to the damage area.

Various types of self-lubrication (e.g., self-replenishing solid and liquid lubrication, reservoirs for liquid lubricant, lubricant supply facilitated by oxidation or a tribochemical reaction, etc.) can also be viewed as surface healing mechanisms.

Theoretical Considerations

From the thermodynamic point of view, self-repair can be viewed as a nonequilibrium self-organization process that leads to increasing orderliness of the material and thus to decreasing entropy. In most schemes of self-repair, the self-healing material is driven away from the thermodynamic equilibrium either by the deterioration process itself or by an external intervention, such as heating. After that, the composite slowly restores thermodynamic equilibrium, and this process of equilibrium restoration drives the healing [5].

In most situations, damage repair at a certain length scale is achieved at the expense of the deterioration at the lower-length scale, making self-repairing materials multi-scale systems. For example, macroscale cracks are healed at the expense of fracturing macrocapsules of the healing agent embedded in the matrix of the material. In the precipitation-induced damage prevention mechanisms, micro-voids are healed by atomic-scale material transport. Multi-scale

organization is typical for biological materials and tissues, it is therefore not surprising that it plays a central role in biomimetic self-repairing materials.

In order to characterize degradation, it is convenient to introduce a so-called “degradation parameter” ξ , to represent, for example, the wear volume, and a corresponding generalized thermodynamic force, Y^{deg} . When a self-healing mechanism is embedded in the system, another generalized coordinate, the healing parameter, ζ , can be introduced, for example, the volume of released healing agent together with the corresponding generalized force, Y^{heal} . The generalized degradation and healing forces are external forces that are applied to the system, and flows are related to the forces by the governing linear Onsager’s equations

$$\begin{aligned} J^{\text{deg}} &= LY^{\text{deg}} + MY^{\text{heal}} \\ J^{\text{heal}} &= NY^{\text{deg}} + HY^{\text{heal}} \end{aligned} \quad (1)$$

where L, M, N, H are corresponding Onsager coefficients [5].

The degradation force Y^{deg} is an externally applied thermodynamic force that results in the degradation. The healing force Y^{heal} is an external thermodynamic force that is applied to the system. In most self-healing mechanisms, the system is placed out of equilibrium and the restoring force emerges, so this restoring force could be identified with Y^{heal} . Since the restoring force is coupled with the degradation parameter ξ by the negative coefficients $N = M$, it also causes degradation decrease or healing.

Measure of Healing

A quantitative measure is required in order to characterize the efficiency of a healing mechanism and to compare different mechanisms. Several parameters have been suggested, such as (1) restored strength divided by the original strength, (2) fracture toughness divided by the original toughness, and (3) creep life of the material (in the case of damage-prevention mechanisms). Most self-repair mechanisms are not

capable to provide 100 % restoration of the original properties of the material.

Cross-References

► [Lotus Effect](#)

References

1. Ghosh, S.K. (ed.): *Self-Healing Materials: Fundamentals, Design Strategies, and Applications*. Wiley, Weinheim (2009)
2. van der Zwaag, S. (ed.): *Self Healing Materials – An Alternative Approach to 20 Centuries of Materials Science*. Springer, Dordrecht (2007)
3. van der Zwaag, S.: Self-healing behaviour in man-made engineering materials: bioinspired but taking into account their intrinsic character. *Phil. Trans. R. Soc. A* **367**, 1689–1704 (2009)
4. Nosonovsky, M., Amano, R., Lucci, J.M., Rohatgi, P.K.: Physical chemistry of self-organization and self-healing in metals. *Phys. Chem. Chem. Phys.* **11**, 9530–9536 (2009)
5. Nosonovsky, M.: Self-organization at the frictional interface for green tribology. *Phil. Trans. R. Soc. A* **368**, 4755–4774 (2010)
6. Nosonovsky, M., Bhushan, B.: Thermodynamics of surface degradation, self-organization, and self-healing for biomimetic surfaces. *Phil. Trans. R. Soc. A* **A367**, 1607–1627 (2009)

Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components

Moon-Ho Ham¹, Ardemis A. Boghossian¹, Jong Hyun Choi² and Michael S. Strano¹

¹Department of Chemical Engineering,
Massachusetts Institute of Technology,
Cambridge, MA, USA

²School of Mechanical Engineering, Purdue
University, West Lafayette, IN, USA

Synonyms

Biomimetic energy conversion devices; Photovoltaic devices; Self-regeneration; Solar cells

Definition

Self-repairing photoelectrochemical cells are synthetic solar energy conversion devices that illustrate principles of self-regeneration and repair through a reversible assembly/disassembly cycle. Motivated by the repair mechanism in living chloroplasts, this cycle can potentially extend the lifetime of a photoelectrochemical cell indefinitely.

Introduction

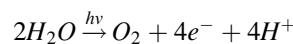
Advancements in nanotechnology toward the synthesis and manipulation of materials have enabled the direct interfacing of synthetic components with biological complexes to create new functions. Direct coupling of synthetic, nanoscale components with biological complexes has been used to study biological systems [1, 2] as well as develop new biomimetic electronics [3–5]. In most recent developments, protein complexes have been interfaced with nanoparticles and single-walled carbon nanotubes (SWCNTs) in the synthesis of biomimetic solar cells [6–10]. This new generation of nano-bio solar cells often relies on a combination of biologically derived components and biologically inspired mechanisms for device fabrication to minimize costs and enhance device efficiencies. As technology progresses in this field, the mechanisms demonstrated by the new devices increasingly mimic processes found in nature. In this work, the hydrophobicity of SWCNTs for the self-assembly of nanoscale components was employed to mimick the self-repair process used by plants [6]. In this sense, a thorough understanding of the natural processes is essential for developing such biomimetic devices.

Photosynthesis: Chloroplast Structure

Natural light-harvesting mechanisms in plants are primarily carried out in the chloroplast, an organelle responsible for the conversion of sunlight into energy for the plant (Fig. 1) [11, 12].

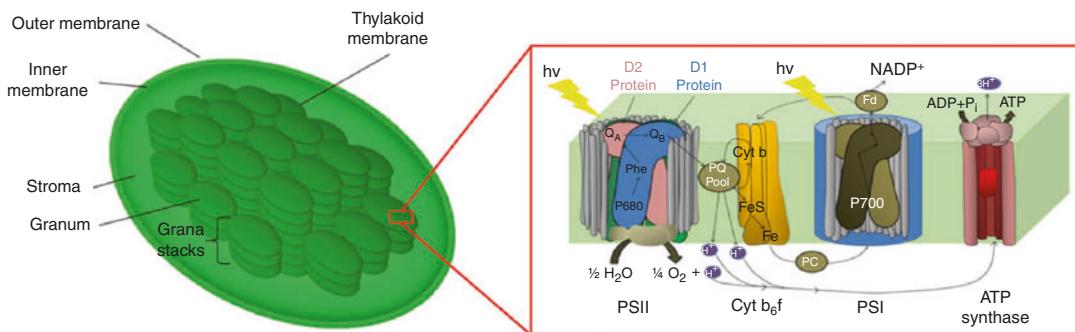
Like most organelles, the chloroplast is surrounded by an outer as well as an inner membrane. A thick, aqueous fluid, the stroma, fills the organelle. The stroma contains genetic materials, including deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and several ribosomes. The most pronounced containment within the chloroplast is the stacks of thylakoid membranes, or grana stacks. Each of the thylakoid membranes contains four key protein complexes used in photosynthesis: photosystem II (PSII), cytochrome b₆f (cyt b₆f), photosystem I (PSI), and adenosine-5'-triphosphate (ATP) synthase. These complexes are responsible for carrying out the first stage of photosynthesis, the light-dependent reactions [11].

The light-dependent reactions are initiated by the absorption of light in PSII. In PSII, light is primarily absorbed by the P680 site at approximately 680 nm. The light-harvesting antennas surrounding by the protein complexes contain an array of pigments and chlorophyll molecules that absorb light at various wavelengths throughout the solar spectrum, enhancing the overall absorbance of light by the plant. The photons absorbed by these antennas are then transferred to the P680 site via resonance energy transfer. Absorption of light at the P680 site excites an electron, transferring it to the pheophytin (Phe) site of the PSII reaction center (RC). The hole remaining in the P680 site is used in the oxidation of water for the production of oxygen and hydrogen.



Meanwhile, the electron is transferred to the Q_A and Q_B sites of PSII, where it is ultimately shuttled via a redox carrier to the next major transmembrane protein complex, cyt b₆f.

Cyt b₆f is responsible for electron transfer between the two major RCs of the membrane, PSI and PSII. The electron transfer between the two centers is coupled to the establishment of a transmembrane proton gradient between the outer stroma and inner thylakoid space. The electron undergoes a series of redox reactions demoting electron energy in exchange for proton transfer



Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components, Fig. 1 Chloroplast structure and function. The chloroplast (*left*) consists of an outer membrane surrounding stacks of thylakoid membranes, or grana. Each thylakoid membrane is embedded with several protein complexes (*right*) including PSII, cyt b₆f, PSI and ATP

from the outer stroma to the inner thylakoid lumen. At the conclusion of this series of redox reactions, the electron is emitted at a lower energy state.

The electron emitted from cyt b₆f is transferred via a redox carrier to the next major protein complex, PSI. Like PSII, PSI is surrounded by a variety of pigment- and chlorophyll-containing antennas. The P700 site of PSI is responsible for light absorption at approximately 700 nm. As before, the surrounding antennas broaden the absorbance of the chloroplast and transfer this absorbed energy to the P700 site via resonance energy transfer. Upon energy absorption, the photoelectric effect is used to excite the electron at the P700, where it is emitted toward the ferredoxin site for pickup by the next major complex. Excited electrons emitted by PSI are then subject to a variety of chemical pathways, one of which is nicotinamide adenine dinucleotide phosphate (NADPH) production. In this reduced state, the NADPH's reducing power is used to power the Calvin cycle, or the second stage of photosynthesis in the light-independent series of reactions, which takes place in the stroma.

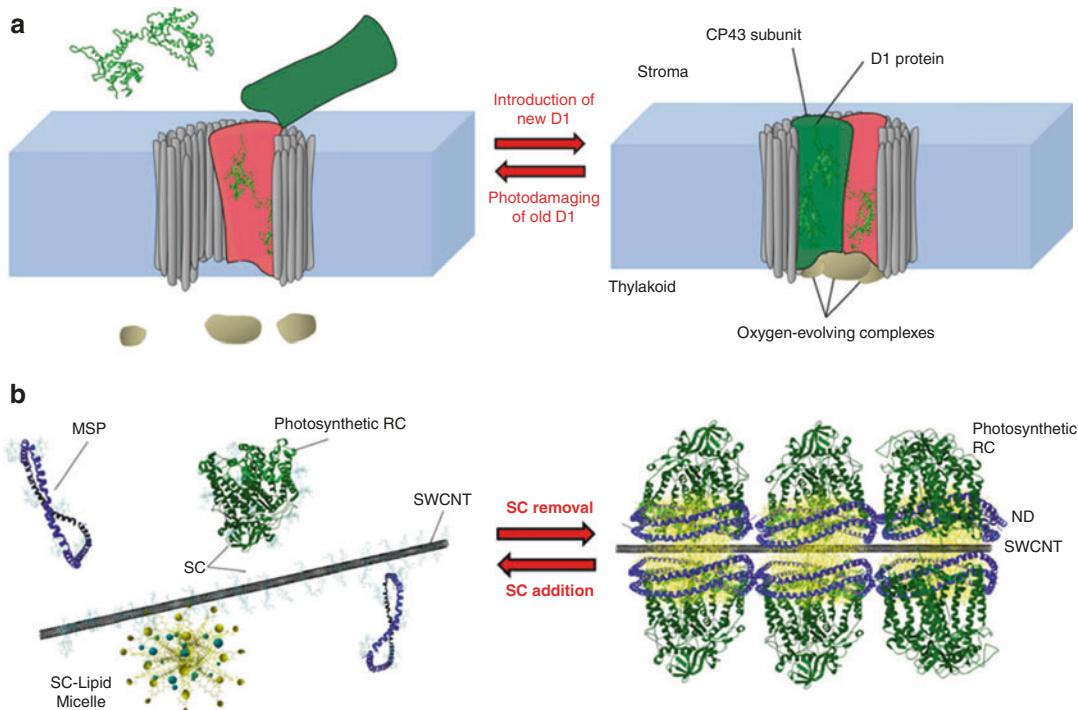
The fourth, and final, major complex embedded within the thylakoid membrane, ATP synthase, is not directly involved in the preceding series of light-dependent reactions. In ATP synthase, a hydrogen ion is extracted from other sites in the membrane, including the water-

synthase. Electron–hole separation occurs in PSI upon light absorption. The hole is used in the oxidation of water into hydrogen and oxygen, while the electron is transmitted through cyt b₆f and PSI. Synthesis of ATP for chemical storage of energy ultimately occurs in ATP synthase

oxidizing site of PSII, to produce ATP. ATP, which is also used in the Calvin cycle mentioned above, is the storage of light-absorbed energy in chemical form. Energy storage and release is carried out by phosphate bond formation and scission, respectively.

Natural Self-repair Cycles in Plants

The aforementioned system of light-dependent reactions relies on a series of embedded protein complexes under nearly continuous illumination that become highly susceptible to photodamage. In particular, the D1 subunit, which contains the P680, Phe, and the Q_B sites of PSII, is vulnerable toward protein damage. To address this, plants have evolved sophisticated mechanisms of self-repair wherein the damaged D1 protein is replaced with a newly synthesized protein (Fig. 2a) [13–15]. The initiation of the self-repair cycle is the damage of the D1 protein, which undergoes a change in conformation and phosphorylation levels when denatured. The dysfunctional protein signals the spontaneous, partial disassembly of the PSII complex. With complex disassembly, the damaged protein is released from the system and diffuses toward the unappressed, stromal region of the membrane. Meanwhile, new D1 protein biosynthesized elsewhere in the cell diffuses toward the appressed



Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components, Fig. 2 Natural and synthetic mechanisms of self-assembly. (a) The self-repair in plants relies on the molecular recognition of components and trapping of meta-stable thermodynamic states. In the disassembled state (*left*), the damage D1 protein is released from the dysfunctional PSII and replaced with a new D1 protein.

region of the membrane. The selective diffusion of the damaged protein toward unappressed regions and new protein toward appressed regions of the membrane is attributed to the varying phosphorylation levels of the two proteins, which vary the degree of protein hydrophilicity. The more hydrophilic protein is driven toward the appressed region of the membrane whereas the more hydrophobic, damaged protein is driven toward the unappressed region. Introduction of the new protein to the disassembled complex triggers the spontaneous reassembly of a functional PSII incorporating the new D1 protein.

The described self-repair cycle consists of transitions between a series of metastable states in an energy-minimized fashion. Aside from the synthesis of a new protein, this cycle requires the input of no additional energy; the reversible

Introduction of the new protein triggers the self-assembly of PSII into a functional state (*right*). (b) The self-assembly of the RC-ND-SWCNT relies on the removal of the surfactant from the system. In the presence of SC, the systems exists in the disassembled state with surfactant dispersion of individual components (*left*). Upon SC removal, the proteins, lipids and nanotubes self-assemble into the complex shown (*right*)

disassembly and reassembly of PSII is completely driven by the formation of thermodynamically and kinetically trapped states. The self-assembly in particular is driven by hydrophobic/hydrophilic interactions and the assortment of diffusivities demonstrated by the proteins on the molecular scale. Synthetic replication of this reversible self-assembly process therefore hinges on the ability to control molecular interactions on the nanoscale.

S

Use of Biological Light-Harvesting Components in Solar Conversion Devices

Biological components such as PSI and photosynthetic RCs have external quantum efficiencies of

almost 100 % and energy yields of approximately 58 %. Recent works have demonstrated that the biological light-harvesting components can be incorporated into optoelectronic devices including photovoltaic devices [6–10].

Jennings and coworkers [10] reported that biomimetically inspired energy conversion devices were fabricated with spinach-derived PSIs which were bound to an Au electrode surface through covalent imine bonds. To increase the surface area of the electrode, the nanoporous leaf-like structure of an Au electrode was employed, which led to the threefold enhancement of the photocurrent due to the increased density of PSI complexes compared to planar electrodes.

Carmeli and coworkers [8] directly bound PSI as a monolayer onto an Au surface. The PSIs were isolated from the cyanobacteria *Synechocystis* sp. PCC6803 and modified at cysteine sites near the P700 for covalent attachment. Kelvin probe force microscopy (KPFM) images show distinct photocurrent generation domains corresponding to PSIs covalently bound to the Au surface with unidirectional orientation. They extended their work to improve the photo-efficiency and apply it to other substrates. Orientated multilayers of PSI complexes were fabricated on the Au surface. In addition, they covalently attached PSI directly to carbon nanotubes and indirectly to GaAs surfaces via a small linker molecule.

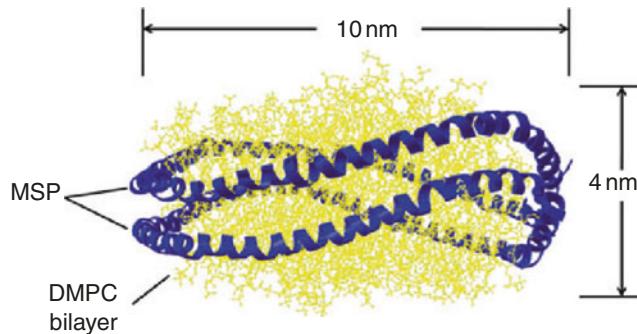
A self-assembled monolayer of RCs, isolated from the purple bacterium *Rhodobacter* (*Rb.*) *sphaeroides*, on an Au electrode with a His tag was incorporated into solid-state photovoltaic cells, as reported by Baldo and coworkers [7]. The photocurrent spectrum of the device matched the absorption spectrum of the RCs, producing internal quantum efficiencies of ~12 %. Lebedev and coworkers [9] used cytochrome *c* as a conductive wire between photosynthetic RCs and the electrode, where the RCs were the same as those in Baldo's group, leading to a remarkable enhancement of the photocurrent. They also used arrayed carbon nanotube electrodes to improve the photo-conversion efficiency, where the RCs were encapsulated inside carbon nanotube arrays. The efficiency was

considerably improved by increasing the number of RCs attached to the electrode surface by about fivefold compared to that obtained with the same proteins when immobilized on a planar graphite (HOPG) electrode.

Until now, most of these efforts have been centered on binding photoactive components onto an electrode surface in a unidirectional orientation to enhance photo-efficiencies. However, even with such an enhancement, these arrangements, like those often found in nature, undergo photo-degradation and protein damage. To address issues of photo-degradation, plants have evolved self-repair mechanisms whereby damaged proteins are replaced with photoactive, newly synthesized proteins. In fact, without this self-assembling mechanism, plants would produce less than 5 % of their typical photosynthetic yields with lifetimes on the order of minutes under intense illumination.

Synthetic Regeneration Cycles in Photoelectrochemical Cells

One approach to mimicking this self-assembly process is to use nanoscaled materials with controllable properties to interface with biologically derived photoactive components to create a photoelectrochemical cell capable of plant-like regeneration. In a recent study [6], this approach was used to develop the first photoelectrochemical cell capable of mimicking key steps in the self-repair cycle. An aqueous solution consisting of SWCNTs, RCs isolated from *R. Sphaeroides*, the phospholipid dimyristoylphosphatidylcholine (DMPC), membrane scaffold proteins (MSPs), and the surfactant sodium cholate (SC) (Fig. 2b, left) is placed within a dialysis bag with a 10,000 molecular weight cutoff and dialyzed against buffer to selectively remove SC from the system. The removal of SC spontaneously triggers the self-assembly of the photoactive complex shown in Fig. 2b (right). This complex contains a series of agglomerates sequentially aligned along the length of a SWCNT. These agglomerates consist of a lipid bilayer disk, or nanodisk (ND) (Fig. 3). In the ND, the DMPC



Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components, Fig. 3 ND structure. NDs consist of DMPC lipids arranged into a bilayer surrounded by two strands

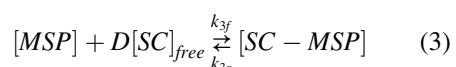
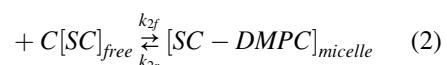
of MSPs. Atomic force microscopy (AFM) measurements reveal that these lipid bilayer disks are approximately 10 nm wide and 4 nm high

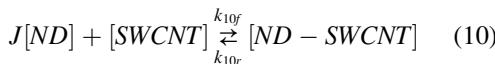
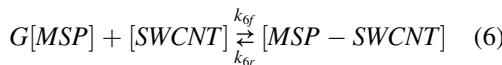
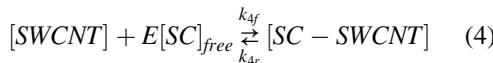
molecules arrange themselves such that their hydrophilic heads face outward toward the aqueous solution and their hydrophobic tails are sandwiched within the bilayer. Wrapped around the circumference of this disk are two strands of MSP, the length of which determines the overall disk diameter. The particular strands of MSP used in this study result in NDs that are approximately 10 nm wide and 4 nm high. These NDs, which align along the length of the nanotube in the assembled state, house photoactive RCs. Although the self-assembly process occurs spontaneously, the RCs are specifically orientated such that their hydrophobic area (the region near the P680 site) faces the SWCNT and their hydrophilic area (the area near the Q_B site) faces outward toward its aqueous surroundings. This self-assembly process is completely reversible, such that the re-addition of surfactant to the solution decomposes the complex back into its initial micellar state, which is monitored using the photoluminescence shift of the nanotubes [6].

Understanding and Quantifying the Self-assembly Process

This spontaneous assembly of nano- and bio-based materials into a precise configuration via chemical signaling alone can be attributed to the comparative scalabilities of the synthetic and biological components and the controlled,

molecular interactions that occur on the nanoscale. The removal of surfactant from the system and, most importantly, from the hydrophobic SWCNT surface triggers the formation of (meta) stable hydrophobic/hydrophilic agglomerates that minimizes SWCNT exposure to its aqueous surroundings. The precise locations and sizes of the hydrophobic/hydrophilic regions in the RC, the thickness of the ND bilayer, and the unidimensionality and length of the SWCNT are all factors that contribute to the precise molecular arrangement of the RCs along the nanotube length. Even the slightest alteration in sizes, diffusivities, and hydrophobicities of the nano-components are enough to perturb the dynamics of the system and even altogether inhibit the formation of such a photoactive complex. To develop a more thorough, quantitative understanding behind the formation of these complexes, the formation of various agglomerates upon surfactant removal was modeled as a series of reactions [16].

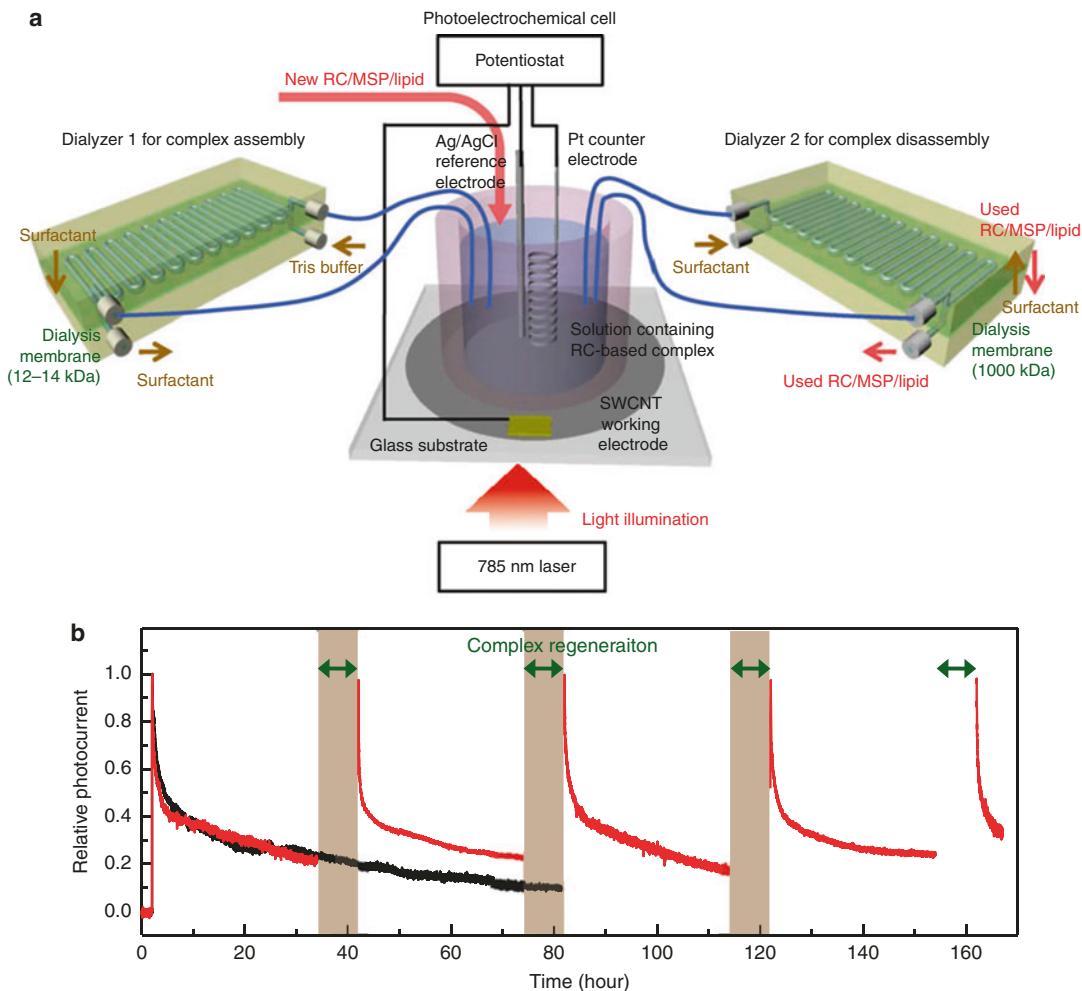




where $[SC]_{free}$ is free sodium cholate concentration in monomeric form, $[SC]_{micelle}$ is sodium cholate concentration in micellar form, $[DMPC]$ is free lipid concentration in monomeric form, $[SC-DMPC]_{micelle}$ is SC-lipid mixed micelle concentration, $[MSP]$ is MSP concentration, $[SC-MSP]$ is SC-suspended MSP concentration, $[SWCNT]$ is free SWCNT concentration, $[SC-SWCNT]$ is SC-suspended SWCNT concentration, $[DMPC-SWCNT]$ is lipid-SWCNT aggregate concentration, $[MSP-SWCNT]$ is protein-lipid aggregate concentration, $[SWCNT_2]$ designates SWCNT bundle concentration, $[SC]_{removed}$ is the concentration of SC that has been removed from the system, $[ND]$ is a ND concentration, and $[ND-SWCNT]$ is ND-SWCNT concentration. In these reactions, k_f is the forward rate constant, k_r is the reverse rate constant, and A-J are stoichiometric coefficients. Fitting this model to experimental data on ND-SWCNT concentration over the course of dialysis, the best-fit rate constants for ND and ND-SWCNT formation were $79 \text{ mM}^{-1}\text{s}^{-1}$ and $5.4 \times 10^2 \text{ mM}^{-1} \text{ s}^{-1}$, respectively. In a diffusion-controlled process, one would expect

the ND-SWCNT rate constant to be smaller than that of the ND, since the bulkier NDs and SWCNTs are expected to have smaller diffusivities than the individual lipid and MSP molecules. However, the reverse is the case, and the calculated ND-SWCNT rate constant is larger than that of the ND, indicating that the system is not under diffusion-controlled conditions. Because of the strongly hydrophobic nature of the SWCNT, the adsorption of the ND is expected to be nearly instantaneous, and in this case, the interaction is expected to be closer to diffusion-controlled conditions than ND formation [16].

The hydrophobic nature of SWCNTs not only induces the formation of ND-SWCNT complexes, but also the formation of other agglomerates, including SWCNT bundles. As discussed above, removal of surfactant from the SWCNTs forces the nanotubes to seek relatively hydrophobic surfaces to shield them from the aqueous surroundings. These surfaces may be lipid-formed agglomerates, such as NDs, or other nanotubes. In a diffusion-controlled dialysis, the structure with the largest diffusion coefficient is expected to diffuse toward the exposed nanotube surface more quickly. Large, bulky nanotubes with lengths on the order of microns demonstrate smaller diffusion coefficients than the more agile NDs; hence, they are expected to diffuse more slowly toward one another than the ND, favoring the faster ND-SWCNT formation. However, although diffusion occurs more slowly, the nanotube bundling is nearly irreversible: once nanotube bundles are formed, their dissociation would require the use of high-energy perturbations via sonication. Over extended periods of time, SWCNTs are expected to occupy their thermodynamically favored state in bundles, where strong, continuous, hydrophobic interactions along the entire length of the nanotube maintain irreversible bundle formation. Slower dialysis conditions would thus favor the formation of kinetically slower, but thermodynamically favored nanotube bundles. Faster dialysis conditions, on the other hand, favor the formation of kinetically faster, but thermodynamically metastable, ND-SWCNT complexes over smaller time scales [6].



Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components, Fig. 4 Regenerating photoelectrochemical cell with self-assembled RC-ND-SWCNT complexes. **(a)** Schematic of the setup of photoelectrochemical systems which consists of a photoelectrochemical cell incorporated to two re-circulating membrane dialyzers. Dialyzers 1 and

2 are used for assembly and disassembly of the complex, respectively. The photodamaged RCs are removed during dialysis for disassembly of the complex and replaced with new ones while SWCNT are retained. **(b)** Temporal photoresponse of the RC-ND-SWCNT complexes with (red line) and without (black line) regeneration. (Reproduced from [6] © 2010 Nature Publishing group)

Photoelectrochemical Measurements of Self-assembled Complexes

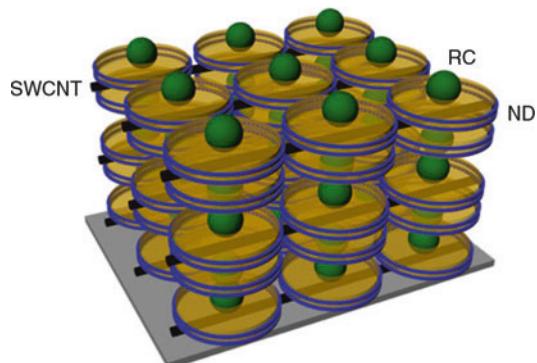
To exemplify the advantages of using the kinetically and thermodynamically favored formation of nanotube-based complexes in dynamic solar cells, the self-assembly was used as a means of regeneration in photoelectrochemical measurements. Photoelectrochemical measurements were

carried out using a three-electrode system: a casted SWCNT working electrode, a Pt counter-electrode, and a Ag/AgCl standard reference electrode. A ferricyanide/ubiquinone dual mediator was used to optimize device efficiency. Illumination of a 700 nM solution of RC-ND-SWCNT yielded a per nanotube external quantum efficiency (EQE) of approximately 40 %. Upon introduction of SC to the illuminated solution, the

photoactive complex disassembles, and no photocurrent is recovered [6].

Although the exact mechanism behind the demonstrated photoactivity of the assembled complex remains unclear, it is hypothesized that electron–hole generation occurs in the RC in a manner similar to that found in nature, where the hole ultimately occupies the P680 site and the electron is shuttled to the Q_B. In nature, the Q_B site contains ubiquinone. Ubiquinone molecularly consists of an electron-withdrawing head and a conjugated tail. In vivo, ubiquinone occupies a thermodynamically favored configuration within the RC wherein the head is docked in the Q_B site and the tail is extended outward toward to the surface of the protein complex. Since photoactivity is only observed in the presence of ubiquinone, it is hypothesized that the ubiquinone docks to the RC in an analogous manner, serving as a means of electron extraction from the Q_B site buried deep within the RC. On the other hand, the P680 site, which is much more easily accessible, faces the SWCNT, which may serve as a hole-conducting wire in this configuration.

Although photoactivity has been demonstrated in the assembled state, it remains to be seen whether reversible self-assembly could be used as a means of solar cell regeneration, as is the case with the plants. To address this, photoelectrochemical measurements were taken over an extended period of continuous illumination, provoking photodamage and diminishing photocurrents (Fig. 4). When the photocurrent approaches approximately 30 % of its initial value, SC is introduced to the system to disassemble the complex, the damaged proteins are replaced with new ones, and surfactant is once more removed from the system to reassemble the complex and recover photocurrent to its initial value. The synthetic regeneration cycle is hence conducted in analogy to the regeneration cycle used by plants, whereby chemical signaling is used to transition between the metastable assembled and disassembled states to replace photodamaged proteins with new ones. Using this regeneration cycle in the synthetic device, photoelectrochemical cell lifetime can be extended indefinitely while increasing efficiency by over 300 % over 168 h.



Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components, Fig. 5 High-density stacks for optimal efficiencies. Arrays of RC-ND-SWCNT can be aligned to maximize concentration of these complexes. Alignment of these complexes relies on recent advancements on SWCNT alignment

Nanotechnology as a Means of Modulating Bio-inspired Devices

The development of the regenerable, self-assembling solar device hinges on the ability to interface SWCNTs with biologically derived components. The precise arrangement of hydrophobic and hydrophilic regimes in biological complexes, such as RCs, is confined to regions that are at most a few nanometers apart. To utilize these variations in non-covalent interactions, one must subject these complexes to materials with properties that can be controlled on the nanoscale. For instance, nanotubes offer unidimensional confinement of hydrophobic surfaces that allows for the one-dimensional alignment of light-harvesting complexes. Confinement of these complexes into high-density, unidirectional arrays allows for device optimization and design. In fact, preliminary findings [6] indicate a linear increase in photo-efficiency with increasing RC-ND-SWCNT complexes, suggesting that maximum efficiency (e.g., 40 %) could be achieved by maximizing concentration. Maximum concentration can be realized through high-density RC-ND-SWCNT stacks (Fig. 5). The key in synthesizing these stacks is controlling the alignment of not only RC-NDs along the nanotube, but also the nanotubes relative to one another. Recent advancements in nanotube

alignment [17–20] have expanded researchers' ability to precisely control device fabrication on the nanoscale, making the fabrication of such high-density stacks feasible.

Cross-References

- [Carbon Nanotubes](#)
- [Hybrid Solar Cells](#)
- [Nanomaterials for Excitonic Solar Cells](#)
- [Nanostructures for Energy](#)
- [Self-Assembly of Nanostructures](#)

References

1. Rosi, N.L., Mirkin, C.A.: Nanostructures in biodiagnostics. *Chem. Rev.* **105**, 1547–1562 (2005)
2. Wong, S.S., Joselevich, E., Woolley, A.T., Cheung, C.L., Lieber, C.M.: Covalently functionalized nanotubes as nanometre-sized probes in chemistry and biology. *Nature* **394**, 52–55 (1998)
3. Oregan, B., Gratzel, M.: Low-cost, high-efficiency solar-cell based on dye-sensitized colloidal TiO₂ films. *Nature* **353**, 737–740 (1991)
4. Colfen, H., Mann, S.: Higher-order organization by mesoscale self-assembly and transformation of hybrid nanostructures. *Angew. Chem. Int. Ed.* **42**, 2350–2365 (2003)
5. Lowe, C.R.: Nanobiotechnology: the fabrication and applications of chemical and biological nanostructures. *Curr. Opin. Struct. Biol.* **10**, 428–434 (2000)
6. Ham, M.H., Choi, J.H., Boghossian, A.A., Jeng, E.S., Graff, R.A., Heller, D.A., Chang, A.C., Mattis, A., Bayburt, T.H., Grinkova, Y.V., Zeiger, A.S., Van Vliet, K.J., Hobbie, E.K., Sligar, S.G., Wright, C.A., Strano, M.S.: Photoelectrochemical complexes for solar energy conversion that chemically and autonomously regenerate. *Nat. Chem.* **2**, 929–936 (2010)
7. Das, R., Kiley, P.J., Segal, M., Norville, J., Yu, A.A., Wang, L.Y., Trammell, S.A., Reddick, L.E., Kumar, R., Stellacci, F., Lebedev, N., Schnur, J., Bruce, B.D., Zhang, S.G., Baldo, M.: Integration of photosynthetic protein molecular complexes in solid-state electronic devices. *Nano Lett.* **4**, 1079–1083 (2004)
8. Frolov, L., Wilner, O., Carmeli, C., Carmeli, I.: Fabrication of oriented multilayers of photosystem I proteins on solid surfaces by auto-metallization. *Adv. Mater.* **20**, 263 (2008)
9. Lebedev, N., Trammell, S.A., Spano, A., Lukashev, E., Griva, I., Schnur, J.: Conductive wiring of immobilized photosynthetic reaction center to electrode by cytochrome c. *J. Am. Chem. Soc.* **128**, 12044–12045 (2006)
10. Ciesielski, P.N., Scott, A.M., Faulkner, C.J., Berron, B.J., Cliffel, D.E., Jennings, G.K.: Functionalized nanoporous gold leaf electrode films for the immobilization of photosystem I. *ACS Nano* **2**, 2465–2472 (2008)
11. Asimov, I.: *Photosynthesis*. Basic Books, New York (1968)
12. Stern, K.R., Bidlack, J.E., Jansky, S.: *Introductory Plant Biology*, 11th edn. McGraw-Hill Higher Education, Boston (2008)
13. Aro, E.M., Virgin, I., Andersson, B.: Photoinhibition of photosystem-2 – inactivation, protein damage and turnover. *Biochim. Biophys. Acta* **1143**, 113–134 (1993)
14. Melis, A.: Photosystem-II damage and repair cycle in chloroplasts: what modulates the rate of photodamage *in vivo*? *Trends Plant Sci.* **4**, 130–135 (1999)
15. Melis, A.: Dynamics of photosynthetic membrane-composition and function. *Biochim. Biophys. Acta* **1058**, 87–106 (1991)
16. Boghossian, A.A., Choi, J.H., Ham, M.H., Strano, M.S.: Dynamic and reversible self-assembly of photoelectrochemical complexes based on lipid bilayer disks, photosynthetic reaction centers, and single-walled carbon nanotubes. *Langmuir* **27**, 1599–1609 (2011)
17. Xie, X.L., Mai, Y.W., Zhou, X.P.: Dispersion and alignment of carbon nanotubes in polymer matrix: a review. *Mater. Sci. Eng. R Rep.* **49**, 89–112 (2005)
18. Vigolo, B., Penicaud, A., Coulon, C., Sauder, C., Pailler, R., Journet, C., Bernier, P., Poulin, P.: Macroscopic fibers and ribbons of oriented carbon nanotubes. *Science* **290**, 1331–1334 (2000)
19. Haggemuelle, R., Gommans, H.H., Rinzler, A.G., Fischer, J.E., Winey, K.I.: Aligned single-wall carbon nanotubes in composites by melt processing methods. *Chem. Phys. Lett.* **330**, 219–225 (2000)
20. Murakami, Y., Chiashi, S., Miyauchi, Y., Hu, M.H., Ogura, M., Okubo, T., Maruyama, S.: Growth of vertically aligned single-walled carbon nanotube films on quartz substrates and their optical anisotropy. *Chem. Phys. Lett.* **385**, 298–303 (2004)

S

SEM

- [Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells](#)

Semiconductor Nanocrystals

- [Quantum Dot Toxicity](#)

Semiconductor Piezoresistance

- ▶ [Piezoresistivity](#)

Semiempirical Methods

- ▶ [Tight-Binding Simulations of Nanowires](#)

Sense Organ

- ▶ [Arthropod Strain Sensors](#)

Sensors

- ▶ [Arthropod Strain Sensors](#)

Shark Denticles

- ▶ [Shark Skin Drag Reduction](#)

Shark Skin Drag Reduction

Amy Lang¹, Maria Laura Habegger² and Philip Motta²

¹Department of Aerospace Engineering and Mechanics, University of Alabama, Tuscaloosa, AL, USA

²Department of Integrative Biology, University of South Florida, Tampa, FL, USA

Synonyms

Riblets, Shark denticles, Shark skin separation control

Definition

The scales, or denticles, on fast-swimming sharks have evolved two mechanisms for controlling the boundary layer flow over the skin surface leading to a reduction in drag. The first, and most widely known and studied, consists of the small streamwise keels covering the surface of the scales also known as riblets which reduce turbulent skin friction drag. The second mechanism is attributed to loosely embedded scales that are located on key regions of the body. When actuated to bristle by the flow, these scales potentially act as a means of controlling flow separation, thereby minimizing pressure drag during swimming maneuvers. Shark scales display a wide variation in geometry both across species while also varying with body location, but on faster swimming sharks, they typically range in size from 180 to 500 μm in crown length.

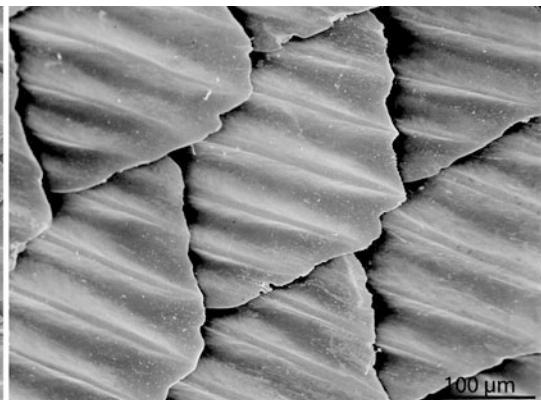
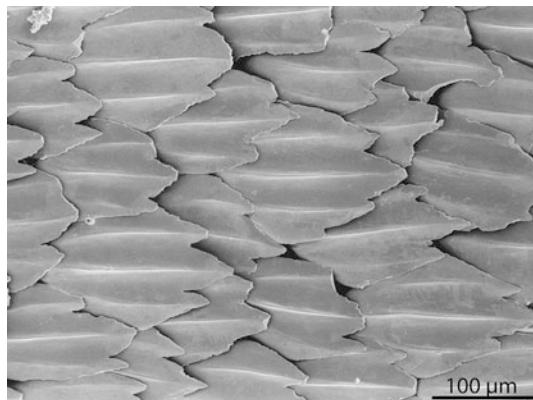
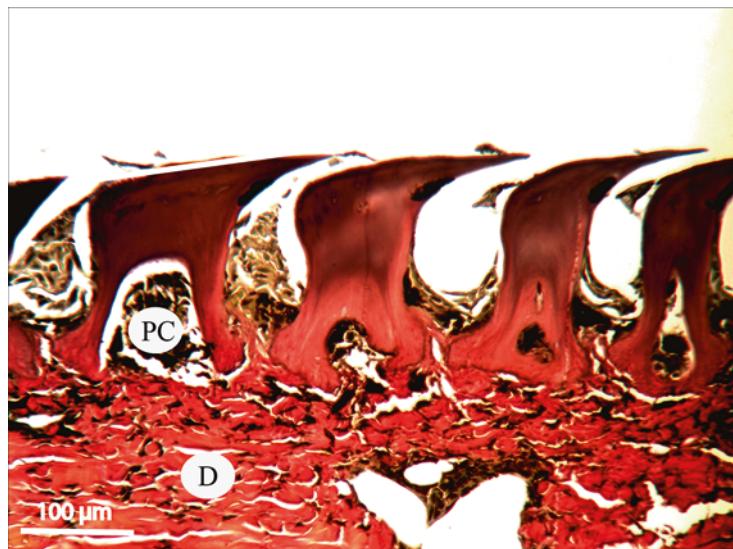
Overview

The Shark Skin

Sharks are covered with minute scales, also known as denticles or placoid scales because of their toothlike nature. The scales have a pulp cavity and a hard enameloid covering and are anchored at the base of the scale to the collagenous layer of the skin known as the stratum laxum (Fig. 1). The interlocking crowns of each scale make up the surface of the shark exposed to the water, and it is on the crown where many species have developed small riblets, or keels, orientated in the streamwise direction of the flow (Fig. 2). A reduction in the length of the base relative to the length of the crown and a change of shape of the base for some species over certain regions of the body appear to be the means by which certain scales have developed the capability to bristle or erect upon flow reversal (Fig. 3; flow would normally pass over the surface from left to right; flow reversal proceeding right to left can induce scale bristling as shown in Fig. 4 with a schematic shown in Fig. 5). The length of the scales is typically fixed for specific regions of the body within a species but differs among regions and

Shark Skin Drag Reduction

Reduction, Fig. 1 Lateral view of sectioned placoid scales of a shortfin mako shark *Isurus oxyrinchus* in the region midway between the leading and trailing edge of the pectoral fin. Note the relatively long scale base relative to the crown length. The pulp cavity (*PC*) of the scale on the left is visible, and the base of the scales is anchored by collagen fibers to the dermis (*D*) of the skin



Shark Skin Drag Reduction, Fig. 2 Scanning electron micrograph (200 \times) of the placoid scales of a shortfin mako shark *Isurus oxyrinchus* (left) and blacktip shark *Carcharhinus limbatus* (right) from the dorsal body wall

species. Similarly, the number of keels per scale is also consistent per location for a species. For instance, on the fast-swimming shortfin mako (*Isurus oxyrinchus*), the flank scales have a crown length of approximately 0.18 mm; each crown typically has three keels, each having a height of 0.012 mm and a spacing of 0.041 mm. The slower swimming blacktip shark (*Carcharhinus limbatus*) has flank scales typically 0.32 mm in length; each crown typically has five keels with a height of 0.029 mm and a spacing of 0.065 mm (Fig. 2).

anterior to the dorsal fin. The mako shark scale has three keels or riblets, whereas the blacktip shark has five. Anterior is to the left

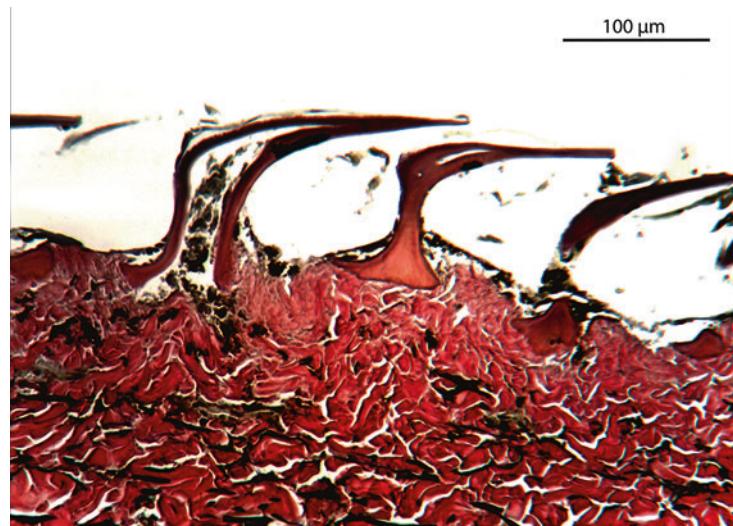
S

Specifications and Fluid Dynamics

A shark swimming through water experiences two major sources of drag due to the viscous resistance of the fluid flow in the direction of motion. It is generally accepted that faster swimming sharks, such as the shortfin mako (*Isurus oxyrinchus*), have the capability to reduce both types of drag through evolutionary adaptations to their denticles. The first source of drag is skin friction drag; though not necessarily the largest contributor to overall drag, it is the one most associated with friction of the flow moving over the body of

Shark Skin Drag Reduction

Reduction, Fig. 3 Lateral view of sectioned placoid scales in the flank region of a shortfin mako shark *Isurus oxyrinchus* in the region midway between the dorsal fin and the pectoral fin. Note the relatively long scale crown relative to the shorter base length

**Shark Skin Drag Reduction**

Reduction, Fig. 4 Side view through the shortfin mako skin (from the flexible flank area) showing scales that have been manually erected. Because of the individual manual erection, not all scales are erected to the same degree. Flow would normally pass over the skin from left to right, and reversed flow, as occurs during separation, is hypothesized to cause bristling as shown

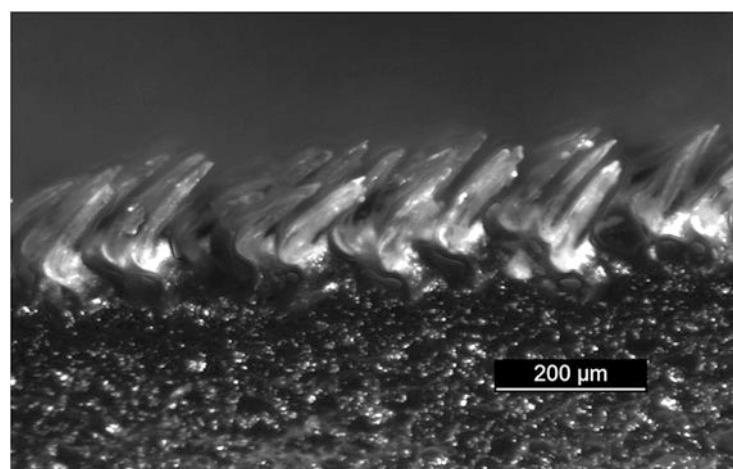
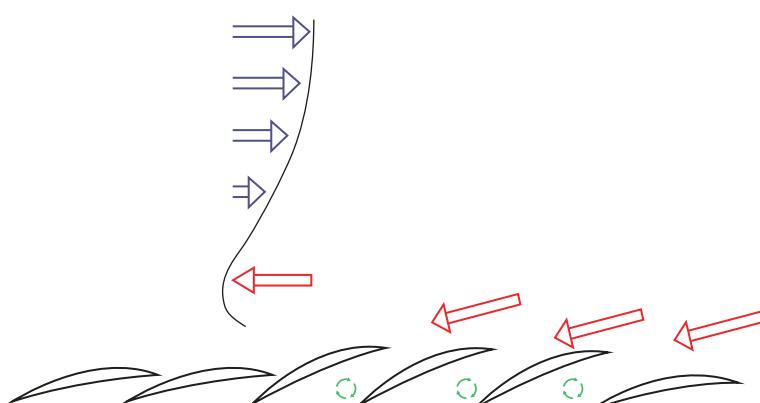
**Shark Skin Drag Reduction**

Fig. 5 Schematic showing forward flow in the boundary layer (blue arrows) subject to an adverse pressure gradient causing flow reversal (red arrows). The reversed flow close to the scales causes localized bristling to occur and the formation of cavity vortices (circular green arrows)



the shark. Skin friction drag is a result of the no-slip boundary condition between the water and the surface of the shark which results in the formation of a boundary layer. Because of the high speeds a shark can achieve (often greater than $U = 10$ m/s, where U is the swimming speed), this leads to a very high Reynolds number ($Re = Ux/v$, where x is length and v is kinematic viscosity) in the boundary layer forming over most portions of the body and indicates the development of turbulent flow over most of the shark's body. For instance, at this speed, the typical transition location when the local $Re = 5 \times 10^5$ occurs at a location just $x = 5$ cm from the nose. While a turbulent boundary layer is less prone to separate from the body, it will have a skin friction magnitude 5–10 times larger than if the flow were to remain laminar. This riblet drag reduction mechanism has demonstrated the potential to reduce the turbulent skin friction drag by approximately 8–10 % in man-made applications [1, 2]. The size and spacing of the riblets are indicative of the speed at which a shark may swim with faster species having closer spaced keels.

The second type of drag, and the most important to be controlled, is due to a difference in pressure around the body and is often referred to as form drag. It is highly dependent on whether the flow remains attached or separates during swimming. Flow separation causes regions of low pressure on the downstream portions of the body leading to an imbalance of the pressure fore and aft and thus a dramatic increase in drag. The first means of decreasing pressure drag is streamlining, which consists of smoothing sharp corners and elongating the body with a tapered downstream portion. As a result, sharks and other marine animals have evolved a very streamlined body shape. Body proportions for fast-swimming sharks such as the shortfin mako are in part determined by its thunniform swimming ability. However, when it undergoes turning maneuvers, the fusiform body will undergo relatively greater body curvature than when swimming in a straight line. Turning may thus induce flow separation, but recent experimental evidence suggests that loosely embedded scales on the flank and behind the gills, in the

region of maximum girth and further downstream, may act as a means of flow control.

Flow separation first involves a reversal of the fluid particles in a thin section adjacent to the surface; this is induced by a region of adverse pressure gradient occurring aft of the point of maximum girth in the streamwise direction. In other words, a suction pressure upstream induces the flow with the least momentum, that closest to the surface, to reverse. Scale bristling likely inhibits the process leading to reversed flow which thereby can control flow separation. Preventing separation will also favorably affect the flow further downstream over the caudal fin which can lead to higher thrust production. While, for a swimming shark, the production of thrust and drag occur simultaneously over the body and are inexorably linked, evidence signifies that sharks have evolved mechanisms within the structure of their skin to reduce drag leading to increases in thrust production and allowing greater maneuverability at high speeds.

Reif, a German biologist, working in the late 1970s is generally considered to be the first to report in literature the hypothesis as to the drag-reducing properties of shark skin [3]. Engineering research began at about the same time in both America [4] and Germany [5] as to the functional aspect of riblet surfaces for reducing skin friction drag. Seminal work into riblet design was completed by Bechert et al. in 1997, which demonstrated a maximum turbulent skin friction drag reduction for man-made, blade-like riblets of 9.9 % [6]. The bristling of shark scales leading to a mechanism for separation control was always hypothesized to function in a manner similar to a man-made technique known as vortex generators utilized since the 1940s [7, 8]. In more recent years, Lang et al. [9] have posited a new mechanism, whereby flow reversal leads to a region of localized scale bristling, leading to a flow-actuated separation control method which can be derived from the shark skin and for which research is ongoing.

Key Research Findings

Since the 1970s when studies began, riblets have become a well-accepted means of reducing

turbulent skin friction drag with an upper limit of reduction of just under 10 %. As previously stated, the most exhaustive testing of various riblet geometries, with comparisons made to other researchers working in the field, was completed in 1997 by Bechert et al. [6]. Previous work by Walsh [4] had focused on a sawtooth geometry, which resulted in a maximum drag reduction of about 5 %. While flow field measurement and visualization were not carried out to fully understand the mechanism behind the drag reduction, an exhaustive series of drag measurement experiments was carried out in a specially designed oil channel facility [6]; these measurements allowed for a determination of the most effectual geometry for man-made riblet applications.

Bernard and Wallace provide a summarization of the basic research that has been performed to map out the key characteristics found in turbulent flows [10]. To understand the mechanism whereby drag is reduced by surface modifications, some basic aspects of a turbulent boundary layer flow field need to be understood. A wall-bounded turbulent flow consists of a layer of vorticity within which fluid located further away from the surface is moving faster than that nearer to the surface due to the no-slip condition, and this results in an overall rotational characteristic of the flow in the clockwise direction for flow moving from left to right where the surface itself is stationary (this is the common reference frame to be used for studying a boundary layer flow). Within this layer, a complex assortment of horseshoe- or hairpin-shaped vortices of various sizes and stages of growth/decay are formed and interact. Scaling of the flow within the boundary layer can be achieved by considering a viscous length scale defined as $\delta_v = v/u^*$, where u^* is the friction velocity. The friction velocity is a function of the shear stress, or skin friction, at the surface (τ) and the fluid density (ρ) such that $u^* = (\tau/\rho)^{1/2}$. The viscous length scale determines the characteristic sizing of the fluid scales found in the boundary layer. For instance, in a given flow with fixed viscosity, v , and at a particular downstream location, x , within a boundary layer consider a variation in free-stream velocity, U . As U is increased, the thickness of the boundary layer at

that location will decrease, and the local average shear stress at that same location will increase. This results in an increase in the friction velocity and thus a decrease in the viscous length scale. The resulting decrease in viscous length scale reduces the characteristic sizing of the vortices forming within the boundary layer.

In the region close to the surface, longitudinal vortices with an axis of rotation in the streamwise direction are found to form and persist. These vortices have a characteristic diameter of approximately $30\delta_v$ and a streamwise length ranging from a few hundred up to 1,000 viscous length scales. When these vortices pair up, a region of low-speed fluid is lifted up from the wall, resulting in the formation of a low-speed streak. It is the instability of these streaks located in a region of high shear that leads to the process known as a turbulent burst and subsequent turbulent sweep. A burst is a sudden ejection of low-speed fluid up into the boundary layer, and a sweep is a sudden injection of high-speed fluid down toward the wall. It is the sweep of this high-momentum fluid onto the surface that results in localized, time-varying patches of high skin friction that are the main causation of increased drag in a turbulent boundary layer. It is the interaction of the shark skin with these components of the boundary layer flow that is essential to controlling the flow.

Results from the work of Bechert et al. [6] plotted the relative change in skin friction ($\Delta\tau/\tau$) versus spacing (s^+), where riblet spacing(s) is nondimensionalized such that $s^+ = s/\delta_v$. They found that the reduction in skin friction increases and reaches a maximum in the vicinity of $s^+ \sim 16$ and thereupon begins to decrease such that larger spacing can actually result in an increase in drag. It is noteworthy that this value corresponds to half the characteristic diameter of the longitudinal vortices forming close to the wall. Thus the riblets sized correctly restrain these near-wall vortices, which for reduced skin friction and reduced viscous length scale will now form at a slightly larger size, and these in turn induce the formation of low-speed streaks [1]. It was also found that a height of the riblets corresponding to half the

spacing ($h = 0.5$ s) gave optimal results. The other key result to garner from these experiments is the variation in maximum decrease with geometry, where a blade-like shape provided the upper limit for skin friction reduction (9.9 %). However, for durability of the surface in real applications, a trapezoidal configuration was tested and found to provide improved performance over the sawtooth geometry. Finally, it is remarkable that the geometry found on shark skin scales, with keels of a scalloped shape closely resembling the trapezoidal shape but with smoothed corners, was also tested by Bechert et al. [6] and found to perform comparably to that of the trapezoidal shape. Many shark species also appear to have the approximate $h = 0.5$ s ratio found to be optimal in experiments. Both of these are indicators that sharks have indeed evolved a scale and riblet geometry for turbulent skin friction reduction; however, the keels may play a role in separation control as well.

Experiments to discern the benefits of shark scale bristling began with Bechert et al. [8]; they used a shark skin replica whereby an overlapping array of individual shark scales was built and tested in their oil tunnel facility. In this man-made model of the shark skin, they meticulously built 800 small replicas of a hammerhead (*Sphyrna zygaena*) shark denticle scaled up 100 times in size (resulting in a crown length of 19 mm and riblet spacing of 4 mm). Each scale was anchored with a compliant spring with variation in stiffness to discern if denticle bristling could result in any additional mechanism to further decrease turbulent skin friction drag. The model was used for cases where the scales laid flat or were wholly bristled at a collective angle of attack. Bristled cases only resulted in increased drag, with higher spring stiffness resulting in higher drag. Flat, aligned scales gave comparable results to riblet surfaces with the only difference being that the greatest drag reduction achieved around $s^+ = 15$ had a value of about 3 %. This decrease in performance was attributed to the construction of the model with small gaps and other imperfections preventing a smoother surface. Thus, if bristling of the scales is to be advantageous to the shark, it must be something that is only activated upon demand and thus concurs

with the postulation that bristling is utilized as a means to control flow separation.

Initial speculation began with the hypothesis that bristled shark scales act as vortex generators [8]. These devices produce streamwise vortices which energize the flow close to the surface. Vortex generators need to be placed at a specific downstream location within a boundary layer for maximum performance and typically upstream of the point of separation [9]. Another method of controlling flow separation, used to date at a more global scale, consists of movable flaps; for airfoil applications, these are placed close to the trailing edge (~10 % of chord length or larger) and have been shown to delay the onset of stall resulting in greater lift [8]. When the flap itself was given a 3D jagged trailing edge, they were also found to act more effectively.

More recent work by the authors [11–15] has investigated the mechanism by which shark skin bristling may lead to the development of a passive, flow-actuated mechanism for separation control. The current working hypothesis is that the passive, flexible scales of the shark work as microflaps to locally control flow separation as needed on crucial regions of the body where it most often occurs during swimming maneuvers. Recent observations on shark skin bristling angles on the shortfin mako (*Isurus oxyrinchus*) indicate that, for this particular species known for its swimming capability, only certain portions of the body have very flexible scales. Bristling capability was measured on dead specimens, and the effect of body pressurization was also considered as a potential bristling mechanism. However, results showed that subcutaneous skin pressurization did not cause scale bristling and had no effect on bristling angles; furthermore, such scales have no muscles attached to them as they lie in the deeper layer of the skin, the dermis. Because the flank scales are very loosely attached to the skin and highly mobile, these observations led us to infer that the scales are most likely bristled by reverse flow actuation. Sixteen body locations [12–14] were considered, six on fins and ten on the body. Scales were manipulated by a fine acupuncture needle and remained bristled once manipulated as shown in Fig. 4.

Scale angles vary with body location, but the most flexible scales are found along the flank of the body extending behind the gills to the tail; here scales are found to be easily flexible with slight manipulation on dead specimens to angles of 50° or greater. Highly flexible scales are also found at the trailing edge of the pectoral fins as compared to the leading edge where there was zero scale flexibility; this indicates the scales may be used to control dynamic stall (unsteady separation) to maintain lift forces on these surfaces during swimming and thereby maintain control. Contragility during swimming, or the ability to change direction quickly and easily, requires low pressure drag as well as high musculature control. These recent findings, herein reported as to variation in scale flexibility, corroborate the hypothesis that the flank region, extending from the location of maximum girth to the tail, is where a shark with a side-to-side swimming motion requires separation control to increase contragility. The scale flexibility appears to be a result of a reduction in size of the scale base anchored in the skin and a change in the shape of the base (as can be seen by the histological data shown in Figs. 1 and 3). The reduction appears to occur in the length of the base relative to its width, where the portion of the base that would pivot up and out of the skin at high bristling angles shows a decrease in length. Thus, the scales with greater bristling angles are less firmly anchored in the anterior to posterior direction and can pivot more freely within the skin.

Cassel et al. [16] describe the process leading to unsteady flow separation, as would occur in a turbulent boundary layer. In the presence of an adverse pressure gradient, the fluid closest to the wall, which also has the lowest momentum, is where flow reversal is first initiated (Fig. 5). This region of fluid moves back upstream and thickens and then subsequently erupts from the surface leading to a patch where the flow is separated from the surface. In a turbulent boundary layer, the fluid with the lowest momentum is that contained in the low-speed streaks, and separation of a streak results in the formation of a horseshoe-shaped structure as observed in computational studies of a separating turbulent boundary layer [17]. Thus for the shark skin, made up of a

staggered array of scales, flow reversal in long, thin patches of fluid may locally bristle the scales, thereby interrupting the flow separation process. This feature of shark skin resulting in a surface with a favored flow direction is likely key to its capability to inhibit flow separation.

Furthermore, previous experiments over a bristled shark skin model established the existence of embedded cavity vortices, axis of rotation in spanwise direction, forming between replicas of the scales [11]. Thus, if flow is induced to form between the scales when bristled, there are two supplementary mechanisms that may help to control the flow. The formation of embedded vortices, similar to those created by golf ball dimples, would allow the flow to pass over the surface with an ensuing partial slip condition, thereby leading to higher momentum adjacent to the skin. Secondly, with a turbulent boundary layer flow forming above the cavities, there may be added momentum exchange whereby high-momentum fluid is induced at a greater rate to move toward the skin and into the cavities. This latter mechanism, resulting in turbulence augmentation [1], is another possible means to enhance the momentum overall in the flow closest to the wall. These three mechanisms may be working in combination to control flow separation over the skin of the shark. Finally, experimental verification that shark skin can control flow separation was recently obtained [15]. In two different experiments using real shark skin specimens, a pectoral fin and hydrofoil covered with flank skin, flow separation was controlled for both laminar and turbulent boundary layer conditions.

Past and Future Applications

The application of drag-reducing techniques to vehicles in both air and water has obvious advantages and at the same time limitations. Typical riblet spacing for both water and air applications, at speeds normally encountered, falls in the range of ~ 0.035 mm. Thin plastic films with adhesive on one side and riblets on the other have been made commercially available; thus a wide array of testing and use on aircraft has already taken place

where drag reduction was documented. However, practical limitations from cost, added weight, and maintenance of the riblet film (particularly with respect to particulates clogging the surface) with only an associated total decrease in drag of 1–3 % have prevented widespread use in most aircraft applications [1].

Separation control cannot only reduce drag but can also lead to increased maneuverability for vehicles. Decreasing drag overall can lead to increased fuel efficiency, payload and range in both military and commercial applications. Flow separation is also an important issue for maintaining use of control surfaces on vehicles (i.e., prevention of stall), which can also have relevance to helicopter rotors and turbine/compressor blades. Passive control mechanisms, including those found on shark skin, have been and will continue to be applied in all these applications.

Cross-References

- ▶ [Biomimetics](#)
- ▶ [BioPatterning](#)
- ▶ [Shark Skin Effect](#)

References

1. Gad-el Hak, M.: *Flow Control: Passive, Active and Reactive Flow Management*. Cambridge University Press, Cambridge, UK (2000)
2. Bhushan, B.: *Shark skin effect*. In: *Encyclopedia of Nanotechnology*. Springer, Berlin (2012)
3. Reif, W.-E.: Protective and hydrodynamic function of the dermal skeleton of elasmobranchs. *Neues Jahrb. Geol. Palaontol. Abh.* **157**, 133–141 (1978)
4. Walsh, M.: Drag characteristics of V-groove and transverse curvature riblets. In: Hough, G.R. (ed.) *Viscous Flow Drag Reduction. Progress in Astronautics and Aeronautics*, vol. 72, pp. 168–184. AIAA, New York (1980)
5. Bechert, D., Hoppe, G., Reif, W.: On the drag reduction of the shark skin. American Institute of Aeronautics and Astronautics (AIAA) Paper No. 85-0546 (1985)
6. Bechert, D., Bruse, M., Hage, W., Van der Hoeven, J., Hoppe, G.: Experiments on drag-reducing surfaces and their optimization with an adjustable geometry. *J. Fluid Mech.* **338**, 59–87 (1997)
7. Bushnell, D., Moore, K.: Drag reduction in nature. *Annu. Rev. Fluid Mech.* **23**, 65–79 (1991)
8. Bechert, D., Bruse, M., Hage, W., Meyer, R.: Fluid mechanics of biological surfaces and their technological application. *Naturwissenschaften* **80**, 157–171 (2000)
9. Lin, J.: Review of research on low-profile vortex generators to control boundary-layer separation. *Prog. Aerosp. Sci.* **38**, 389–420 (2002)
10. Bernard, P., Wallace, J.: *Turbulent Flow: Analysis, Measurement, and Prediction*. Wiley, Hoboken (2002)
11. Lang, A., Motta, P., Hidalgo, P., Westcott, M.: Bristled shark skin: a microgeometry for boundary layer control? *Bioinspir. Biomim.* **3**, 046005 (2008)
12. Lang, A., Habegger, M., Motta, P.: Shark skin boundary layer control. In: *Proceedings of the IMA Workshop “Natural Locomotion in Fluids and on Surfaces: Swimming, Flying, and Sliding,” 1–5 June 2010. IMA Volumes in Mathematics and its Applications* (2012)
13. Lang, A., Motta, P., Habegger, M., Hueter, R., Afroz, F.: Shark skin separation control mechanisms. *Mar. Technol. Soc. J.* **45**(4), 208–215 (2011)
14. Motta, P., Habegger, M., Lang, A., Hueter, R., Davis, J.: Scale morphology and flexibility in the shortfin mako *Isurus oxyrinchus* and the blacktip shark *Carcharhinus limbatus*. *J. Morphol.* **273**(10), 1096–1110 (2012)
15. Lang, A., Bradshaw, M., Smith, J., Motta, P., Habegger, M., Hueter, R.: Movable shark scales act as a passive dynamic micro-roughness to control flow separation. *Bioinspir. Biomim.* **9**, 036017 (2014)
16. Cassel, K., Smith, F., Walker, J.: The onset of instability in unsteady boundary-layer separation. *J. Fluid Mech.* **315**, 223–256 (1996)
17. Na, Y., Moin, P.: Direct numerical simulation of a separated turbulent boundary layer. *J. Fluid Mech.* **374**, 379–405 (1998)

Shark Skin Effect

Bharat Bhushan
Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

S

Synonyms

[Low fluid drag surface](#)

Definition

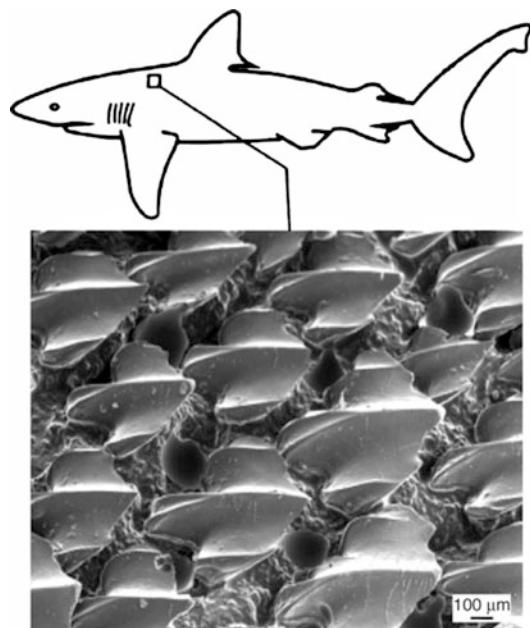
The shark skin effect is the reduction of the fluid drag during swimming at fast speeds and

protection of its surface against biofouling. The presence of surface microstructure on the skin surface is responsible for this effect.

Overview

Many structures, materials, and surfaces found in nature can be exploited for commercial applications. As an example, nature has created ways of reducing drag in fluid flow, evident in the efficient movement of fish, dolphins, and sharks [10, 14]. The mucus secreted by fish causes a reduction in drag as they move through water, and also protects the fish from abrasion by making the fish slide across objects rather than scrape, and it makes the surface of the fish difficult for microscopic organisms to adhere to [28]. It has been known for many years that by adding as little as a few 100 parts per million guar, a naturally occurring polymer, friction in pipe flow can be reduced by up to two thirds. Other synthetic polymers provide an even larger benefit [18]. The compliant skin of the dolphin has also been studied for drag-reducing properties. By responding to the pressure fluctuations across the surface, a compliant material on the surface of an object in a fluid flow has been shown to be beneficial. Studies have reported 7 % drag reduction [12].

Another set of aquatic animals which possess multipurpose skin is fast swimming sharks [14]. The skin of fast swimming sharks reduces the drag experienced by sharks as they swim through water and protects against biofouling. The tiny scales covering the skin of fast swimming sharks, known as dermal denticles (skin teeth), are shaped like small riblets and aligned in the direction of fluid flow (Fig. 1). Shark skin-inspired riblets have been shown to provide a drag reduction benefit up to 9.9 % [5]. The spacing between these dermal denticles is such that the riblets may not be very effective against very small (micro-)organisms – they probably work best against larger organisms such as mussels, algae, and barnacles. Prevention of undesirable accumulation of microorganisms protects the surface against biofouling. Slower sharks are



Shark Skin Effect, Fig. 1 Scale patterns on fast-swimming sharks (*Squalus acanthias*) [19]

covered in dermal denticles as well, but these are not shaped like riblets and do not provide much drag reduction benefits.

The effect of riblet structures on the behavior of fluid drag, as well as the optimization of their morphology, is the focus of this entry.

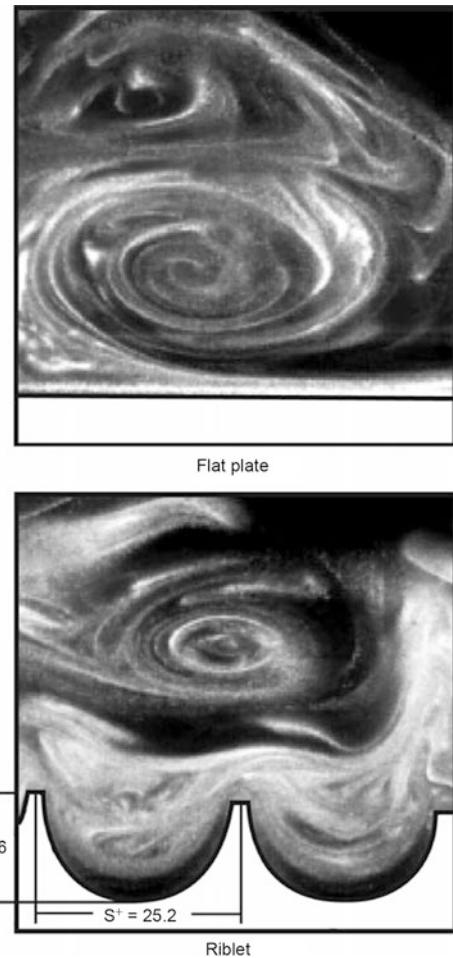
Mechanisms of Fluid Drag and Role of Riblets in Drag Reduction

Fluid drag comes in several forms, the most basic of which are pressure drag and friction drag [14]. Pressure drag is the drag associated with the energy required to move fluid out from in front of an object in the flow, and then back in place behind the object. Much of the drag associated with walking through water is pressure drag, as the water directly in front of a body must be moved out and around the body before the body can move forward. The magnitude of pressure drag can be reduced by creating streamlined shapes. Friction or viscous drag is caused by the interactions between the fluid and a surface parallel to the flow, as well as the attraction between

molecules of the fluid. Friction drag is similar to the motion of a deck of cards sliding across a table. Fluids of higher viscosity – the attraction between molecules – have higher apparent friction between fluid layers, which increases the thickness of the fluid layer distorted by an object in a fluid flow. An increase in drag occurs as fluid velocity increases. The drag on an object is in fact a measure of the energy required to transfer momentum between the fluid and the object to create a velocity gradient in the fluid layer between the object and undisturbed fluid away from the object's surface.

The above discussion of friction drag assumes all neighboring fluid molecules move in the same relative direction and momentum transfer occurs between layers of fluid flowing at different velocities. Fully developed turbulent flow is commonly said to exhibit complete randomness in its velocity distribution, but distinct regions exist within fully developed turbulent flow that exhibit different patterns and flow characteristics [20]. As these vortices rotate and flow along the surface, they naturally translate across the surface in the cross-flow direction. The interaction between the vortices and the surface, as well as between neighboring vortices that collide during translation, initiates bursting motions where vortices are rapidly ejected from the surface and into the outer boundary layers. As vortices are ejected, they tangle with other vortices and twist such that transient velocity vectors in the cross-stream direction can become as large as those in the average flow direction [20]. The translation, bursting of vortices out of the viscous sublayer, and chaotic flow in the outer layers of the turbulent boundary layer flow are all forms of momentum transfer and are large factors in fluid drag. Reducing the bursting behavior of the stream-wise vortices is a critical goal of drag reduction, as the drag reduction possibilities presented by this are sizable.

The vortices have been visualized using flow visualization techniques to capture cross-sectional images, shown in Fig. 2, of the stream-wise vortex formations above both flat-plate and riblet surfaces [26]. The average cross-stream wavelength of these high- and low-speed streaks, the added



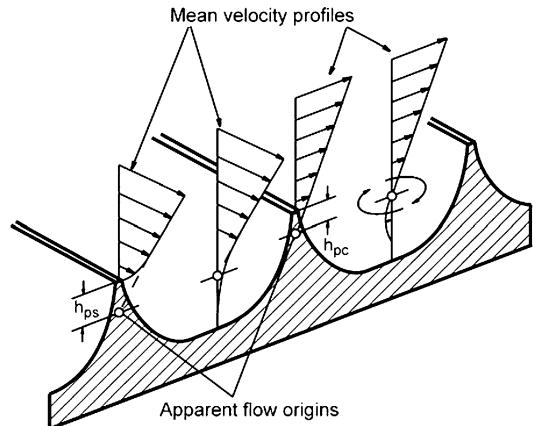
Shark Skin Effect, Fig. 2 Turbulent flow visualization of stream-wise vortices in a vertical cross section over flat-plate and riblet surfaces (Adapted from Ref. [26])

widths of one high-speed streak and one low-speed streak, is equal to the added diameters of two neighboring vortices and has been measured at 70–100 wall units [6, 20, 34]. This corresponds to a vortex diameter of 35–50 wall units. The flow visualizations in Fig. 2 show vortex cross sections and relative length scales, demonstrating vortex diameters smaller than 40 wall units [26]. (As flow properties change, the dimensions of the turbulent flow structures change as well. As such, it is useful to use non-dimensional length values to better compare studies performed in different flow conditions. Dimensionless wall units, marked $^+$, are used for all length scales,

which are calculated by multiplying the dimensional length by V_τ/v . For example, $s^+ = sV_\tau/v$, where s^+ is the non-dimensional riblet spacing, s is the dimensional riblet spacing, v is the kinematic viscosity, and $V_\tau = (\tau_0/\rho)^{0.5}$ is the wall stress velocity, for which ρ is the fluid density and τ_0 is the wall shear stress. Wall shear stress can be estimated for round pipe flow using the equation $\tau_0 = 0.03955v^{1/4}\rho V^{7/4}d^{-1/4}$, where V is the average flow velocity and d is the hydraulic diameter. For flow in rectangular pipes, the equation for hydraulic diameter $d = 4A/c$ can be applied, where A is the cross-sectional area and c is the wetted perimeter.)

The small riblets that cover the skin of fast swimming sharks work by impeding the cross-stream translation of the stream-wise vortices in the viscous sublayer. As vortices form above a riblet surface, they remain above the riblets, interacting with the tips only and rarely causing any high-velocity flow in the valleys of the riblets. Since the higher-velocity vortices interact only with a small surface area at the riblet tips, only this localized area experiences high-shear stresses. The low velocity fluid flow in the valleys of the riblets produces very low shear stresses across the majority of the surface of the riblet. By keeping the vortices above the riblet tips, the cross-stream velocity fluctuations inside the riblet valleys are much lower than the cross-stream velocity fluctuations above a flat plate [26]. This difference in cross-stream velocity fluctuations is the evidence of a reduction in shear stress and momentum transfer near the surface, which compensates the potentially drag-increasing effect of a larger surface area. Though the vortices remain above the riblet tips, some secondary vortex formations do occur that enter the riblet valleys transiently. The flow velocities of these transient secondary vortices are such that the increase in shear stress caused by their interaction with the surface of the riblet valleys is small.

Protruding into the flow without greatly increasing fluid drag allows the riblets to interact with the vortices to reduce the cross-stream translation and related effects. As the riblets protrude into the flow field, they raise the effective flow origin by some distance. The amount by which the height of the



Shark Skin Effect, Fig. 3 Schematic representation of the mean velocity profiles and effective protrusion heights for flow in both the stream-wise direction, h_{ps} , and in the cross-flow direction, h_{pc} (Adapted from Ref. [5])

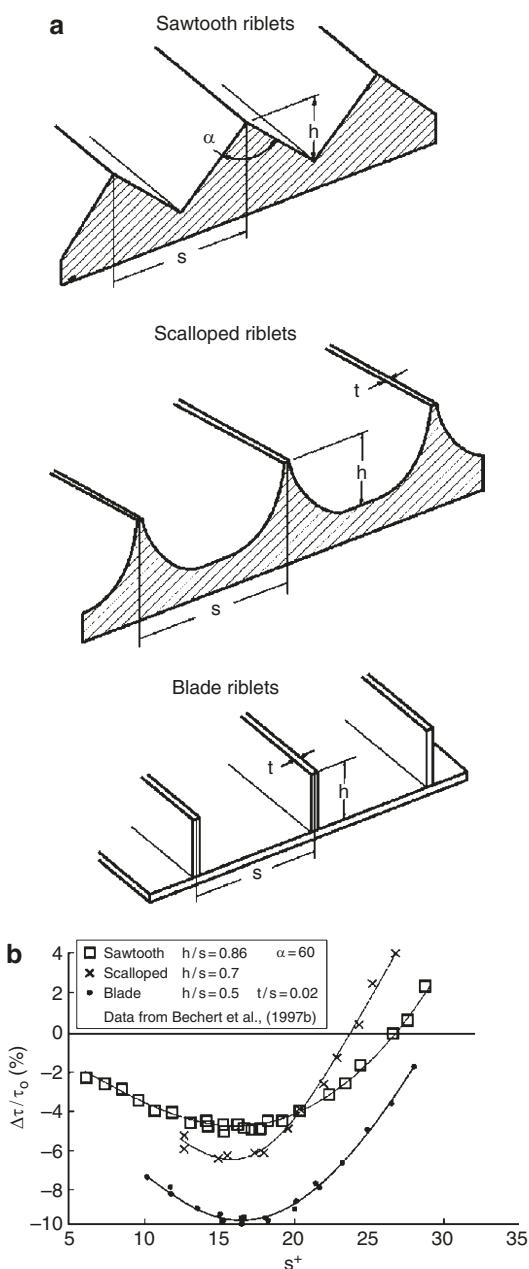
riblets is greater than the apparent vertical shift of the flow origin is referred to as the effective protrusion height. By calculating the average stream-wise velocity in laminar flow at heights over riblet surfaces and comparing them to the average stream-wise velocities in laminar flow at heights over a flat plate, the effective stream-wise protrusion height, h_{ps} , is found for laminar flow. The effective cross-stream protrusion height, h_{pc} , is similarly found for laminar flow by comparing the cross-stream velocities over a riblet surface to those over a flat plate. A schematic of stream-wise and cross-stream flow velocity profiles and effective protrusion heights is shown in Fig. 3. The difference between the vertical shifts in the stream-wise and cross-stream origin, $\Delta h = h_{ps} - h_{pc}$, for any riblet geometry has been proposed to be the degree to which that riblet geometry will reduce vortex translation for low Reynolds number (Re) flows [5]. As Re increases, the degree to which increased surface area affects the overall fluid drag increases, and the drag reduction correlation to the laminar flow theories deteriorates.

Optimization of Riblet Geometry

The cross-sectional shape of riblets on fast swimming sharks varies greatly, even at different locations on the same shark. Many types of riblets

have been studied, the shapes of which have been chosen for several reasons [14]. Rilet shapes have been chosen for their similarity to natural rilets, for their ease of fabrication, and for purposes of drag reduction optimization. Two-dimensional (2D) rilets, which have a continuous extrusion of a simple cross section in the stream-wise direction, have been most extensively characterized. The most thorough characterization has been completed for symmetrical 2D rilets with sawtooth, scalloped, and blade cross sections as shown in Fig. 4a [1, 2, 4, 6, 7, 29–32, 35, 36]. Alternative rilet geometries have, in general, shown no increased benefit. These rilets, including asymmetrical rilets, hierarchical rilets, and rilets with rounded or notched peaks have been studied in detail and do not improve upon the benefit of standard rilet geometries [29, 30, 32]. Other 2D rilet shapes which have been studied include alternating brother-sister-type rilets [4] and hierarchical rilets with small rilets on top of larger rilets [35]. Three-dimensional (3D) rilets, which include segmented 2D rilets as well as shark skin moldings and replicas have also been studied. Rilet types characterized include aligned segmented-blade rilets [35], offset segmented-blade rilets [6], offset-3D blade rilets [6], and 3D shark skin replicas [7, 19, 24].

Most studies are done by changing the non-dimensionalized spacing, s^+ , by varying only fluid velocity and collecting shear stress data from a shear stress balance in a wind tunnel or fluid flow channel. The use of non-dimensional characteristic dimensions for rilet studies, namely, nondimensional spacing, s^+ , is important for comparison between studies performed under different flow conditions. Non-dimensionalization accounts for the change in size of flow structures like vortex diameter, which is the critical value to which rilets must be matched. When comparing the optimal drag reduction geometries for sawtooth, scalloped, and blade rilets shown in Fig. 4b, it is clear that blade rilets provide the highest level of drag reduction, scalloped rilets provide the second most, and sawtooth rilets provide the least benefit [14]. A summary of comparison features for sawtooth, scalloped, and blade rilets is presented in Table 1.



Shark Skin Effect, Fig. 4 (a) Schematic representation of rilet dimensions and (b) drag reduction comparison for sawtooth, scalloped, and blade rilets (Adapted from Ref. [5])

In general, it can be seen in Fig. 4b that each type of rilet is most beneficial near $s^+ \sim 15$, which is between 1/3 and 1/2, the width of the stream-wise vortices. Larger s^+ will cause vortices to begin

Shark Skin Effect, Table 1 Summary and comparison of optimum riblet geometry for various riblet shapes [14]

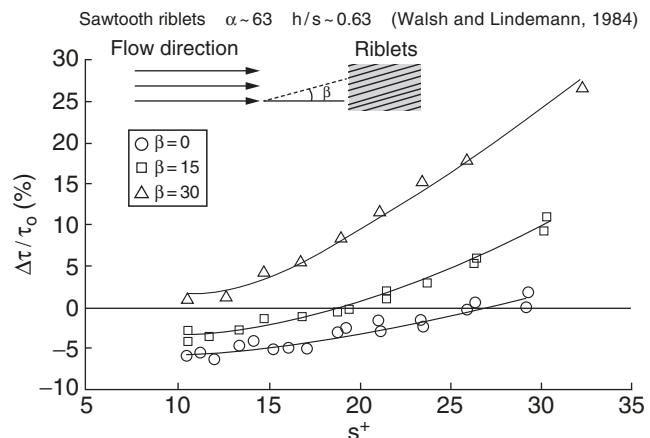
Riblet shape	Relative rank ^a	Maximum drag reduction (%) ^b	Optimum geometry ^b	Comments
Sawtooth	3	5	$h/s \sim 1, \alpha \sim 60^\circ$	Most durable
Scalloped	2	6.5	$h/s \sim 0.7$	
Blade	1	9.9	$h/s \sim 0.5$	Drag reduction increases as riblet thickness, t decreases. Durability is an issue.

^a1 corresponds to greatest drag reduction

^bBased on published data in Ref. [5]

Shark Skin Effect,

Fig. 5 Drag reduction dependence on yaw angle, β , of sawtooth riblets in free stream for $h/s \approx 0.62$ (Adapted from Ref. [32])



falling into the gap between the riblets, which increases the shear stress at the surface between riblets. As s^+ decreases below optimum, the overall size of the riblets decreases to a point below which they cannot adequately impede vortex translation.

An additional concern to the application of riblets is the sensitivity of drag reduction to yaw angle, the angle between the average flow direction and the riblet orientation. Figure 5 shows the effects of yaw angle on riblet performance for flow over sawtooth riblets. Yaw angle has a deleterious effect on the drag reduction benefits of riblet surfaces. Riblet surfaces become drag inducing above $\beta = 30^\circ$, but small drag reductions can still be seen up to $\beta = 15^\circ$ [32].

Riblets on shark skin exist in short segments and groups, not as continuous structures. Riblets with 3D features have been created to better approximate the performance of actual shark skin and to determine if there are methods of drag reduction not yet understood from 2D riblet

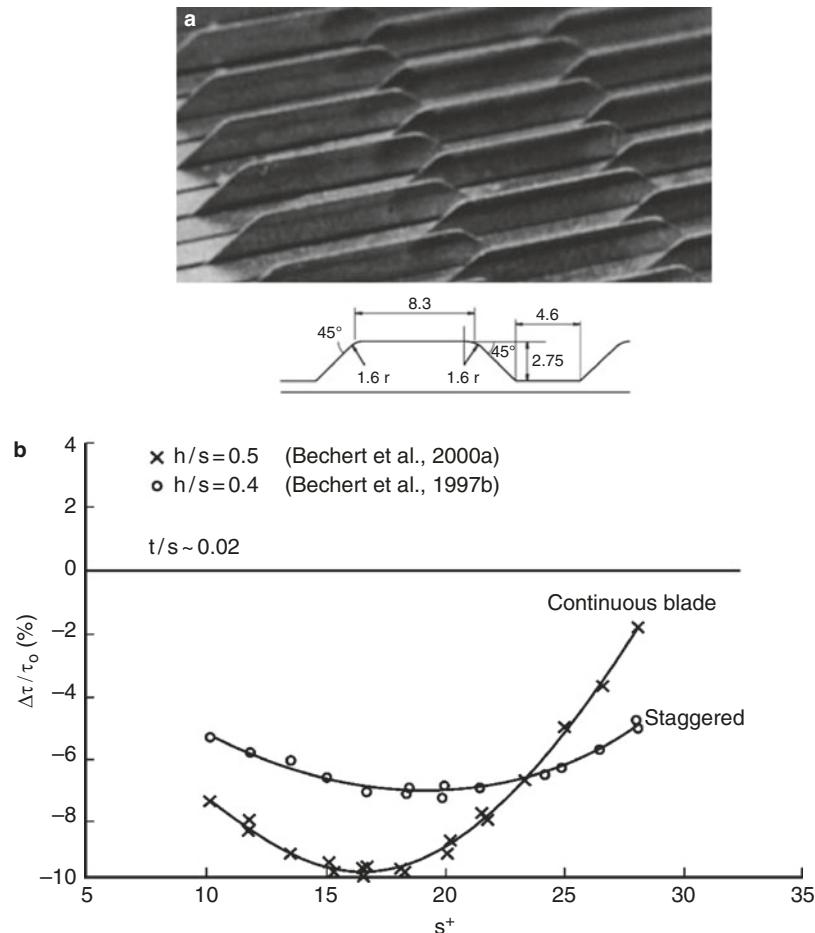
studies. Studies have explored the effects of compound riblet structures and 3D riblets comprised of aligned, segmented-blade riblets [36]. No improvement in net drag reduction was realized when compared to corresponding performance of continuous riblet geometries. More recently, experiments with similarly shaped segmented-blade riblets at spacing s with a matching set of segmented-blade riblets staggered between each row of blades at a spacing of $s/2$ from either side have been performed [5]. A schematic and image of staggered trapezoidal blade riblets is shown in Fig. 6a. Using these and other staggered riblets, Bechert et al. [5] hoped to achieve the same vortex elevation and anti-translational effects of continuous riblets with less effect on the flow origin. Experimental data comparing the largest drag reduction achieved with staggered segmented-blade riblets to optimum continuous blade riblets can be seen in Fig. 6b. Again, no net benefit in drag reduction was achieved, and after comparison of data, the conclusion was made that it is

Shark Skin Effect,

Fig. 6 Comparison of drag reduction over optimum continuous blade riblets with optimum segmented trapezoidal blade riblets. (a) Segmented riblets were staggered as shown.

Spacing between offset rows is $s/2$, while spacing between corresponding rows is s . (b) Optimal hour to second ratio for staggered blade riblets is 0.4.

Staggered blade riblets provide less drag reduction benefit than continuous blade riblets (Adapted from Ref. [5, 6])



unlikely that 3D riblets comprised of segmented 2D riblets will greatly outperform continuous 2D riblets.

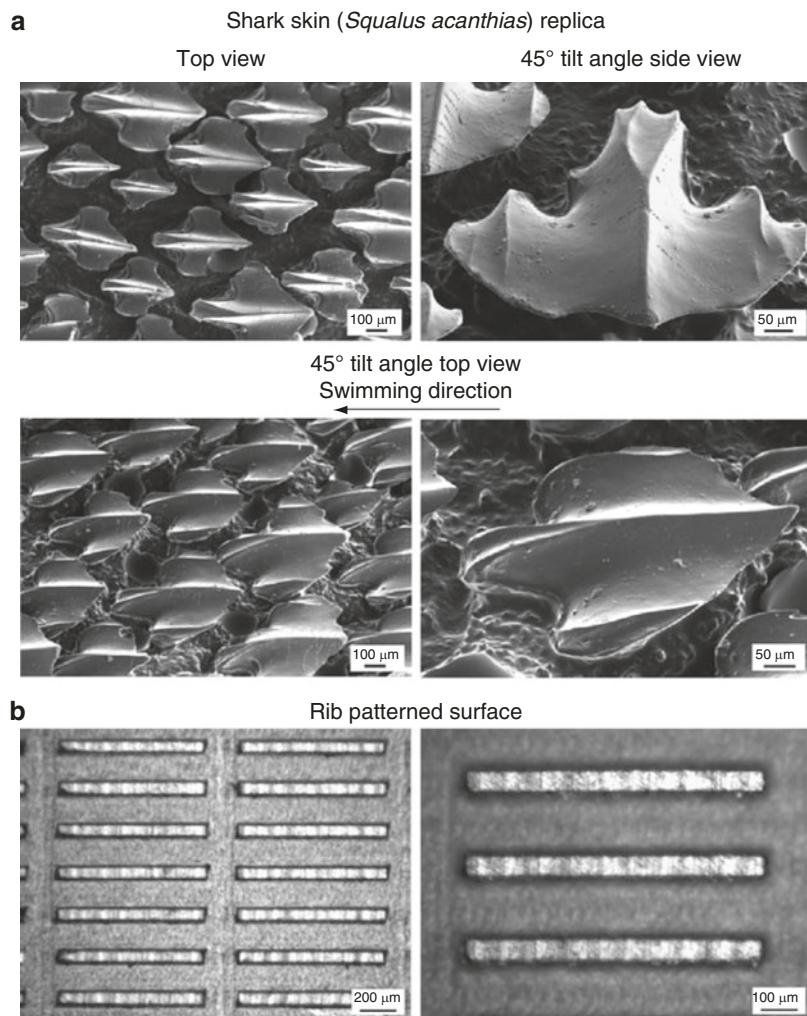
The scales have been molded in epoxy resin from the skin of the Spiny Dogfish (*Squalus acanthias*) by Jung and Bhushan [19], as shown in Fig. 7a. A decrease in pressure drop corresponding to a decrease in fluid drag was achieved as compared to a smooth surface in a rectangular flow cell experiment, as shown in Fig. 8a. Pressure drop from inlet to outlet of a pipe is a measure of drag with large pressure drop occurring as a result of high drag. Aligned riblets were fabricated on acrylic by using micromachining (Fig. 7b), and a minimal decrease in pressure drop was realized as compared to a smooth acrylic test section (Fig. 8b) [19].

Riblet Fabrication and Applications

Riblets have been fabricated for studies and large-scale applications [14]. Typical microscale manufacturing techniques are ill-fitted for large-scale application due to the associated costs. Even for studies, most researchers have opted for traditional milling or molding methods over the microfabrication techniques used in the microtechnology industry. Though non-dimensional units allow for comparison between flow fields of different fluids and at different conditions, the accurate microscale manufacture of riblets for experimentation has been a field of study in its own right. The largest difficulty in optimizing riblet geometries has been the fabrication of riblet series with incremental changes in characteristic dimension. Riblets used in airflow require

Shark Skin Effect,

Fig. 7 (a) Scanning electron microscope (SEM) micrographs of shark skin replica patterned in epoxy, and (b) segmented bade-style riblets fabricated from acrylic [19]

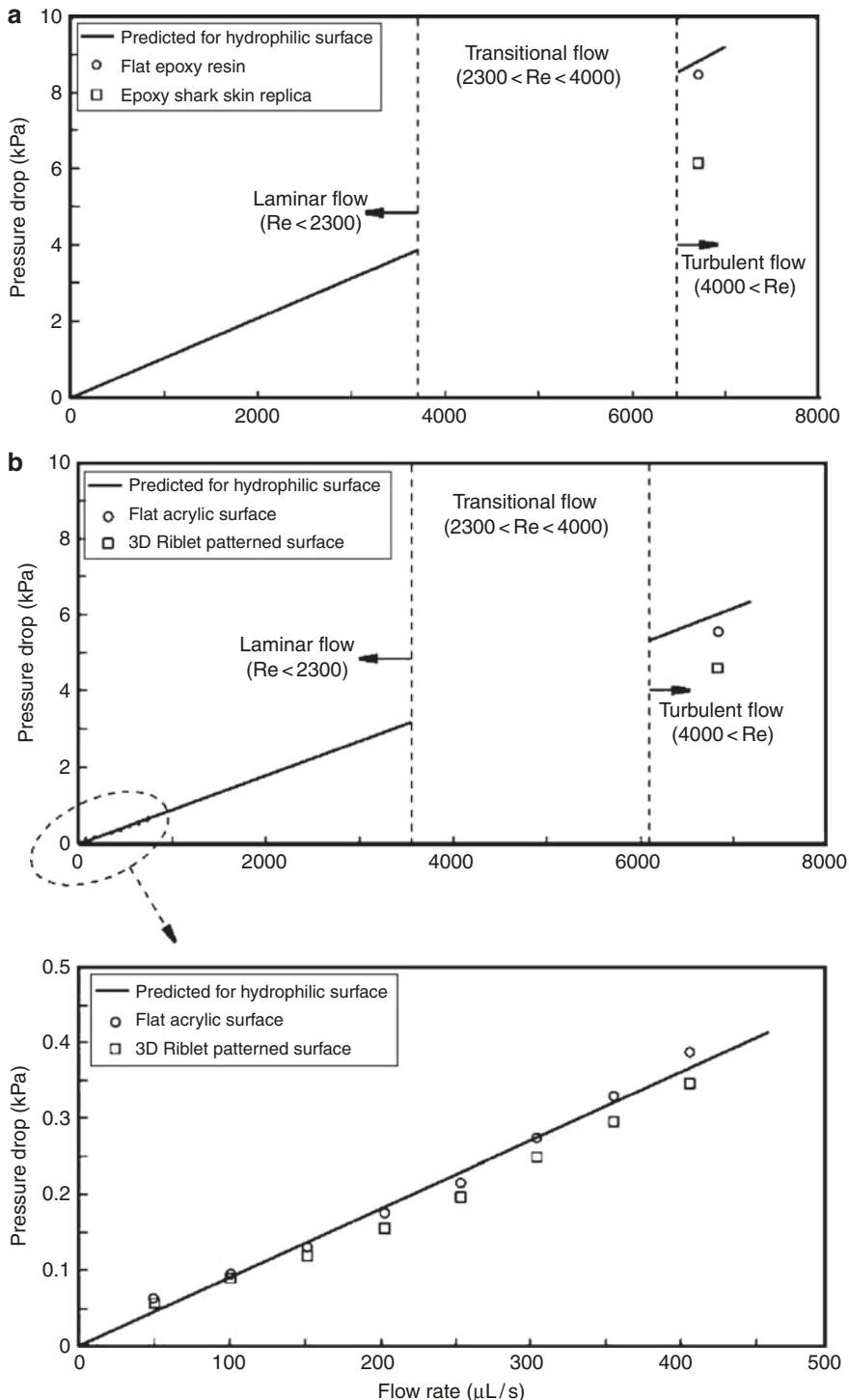


spacings at or below 1 mm due to the low viscosity of air and the high speed at which wind tunnels must operate to create accurately measurable shear stresses on a test surface. Conversely, studies in an oil channel have been carried out in flow that is both highly viscous and slower moving. This allows for riblets to be made with spacings in the 3–10 mm range [3].

Commercial and experimental application of riblets outside wind tunnels and test stands is also limited by the high costs for less-than-optimal riblet performance. Application of riblets on a large scale has been done for several studies as well as for competition and retail purposes. Sawtooth riblets on vinyl films produced by 3 M riblets have been applied on surfaces ranging from

boat hulls to airplanes. Racing swimsuits produced by Speedo and others also employ a riblet pattern on the surface to reduce drag during the streamline portion of each lap of a race [23]. Additionally, a novel surface-scratching technique has been applied to the inside surface of pipelines to create a faux-riblet surface [33].

Beginning in the mid-1980s, vinyl film sawtooth riblets have been applied to boat hulls for racing. Both an Olympic rowing boat and an America's Cup sailing yacht have been covered with riblets during competition. Because skin friction of an airplane accounts for as much as 48 % of total drag, vinyl film riblets have also been applied to test planes of both Boeing and Airbus. These films have not seen use on standard commercial



Shark Skin Effect, Fig. 8 (a) Comparison of pressure drop in rectangular pipe flow over flat epoxy surface with shark skin replica surface. (b) Comparison of pressure drop in rectangular pipe flow over flat acrylic surface and

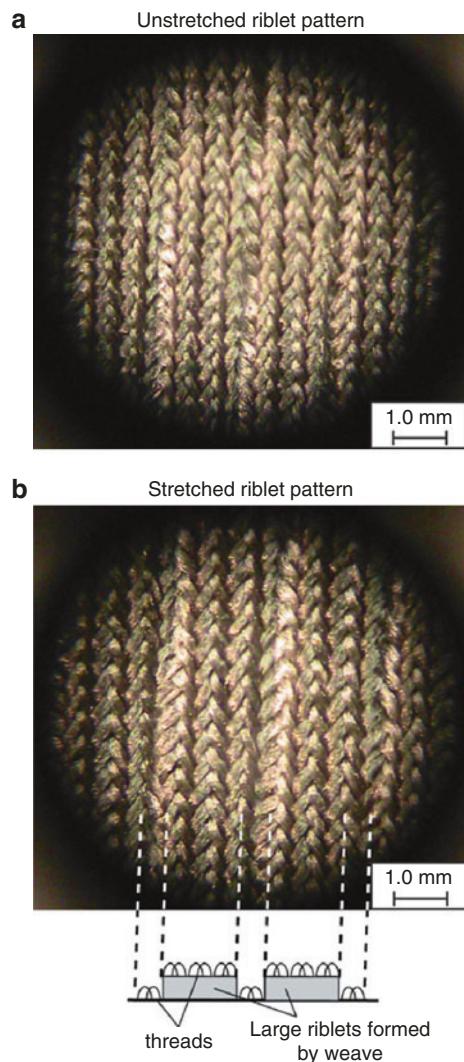
segmented blade riblets. Data are compared with the predicted pressure drop function for a hydrophilic surface (Adapted from Ref. [19])

flights yet, but the benefits seen in testing should not go unmentioned. Application of riblets to an airplane requires that several concessions are made. Several locations that would be covered by riblets must be left uncovered due to environmental factors; windows are not covered for the sake of visibility; several locations where dust and debris contact the airplane during flight are left bare because the riblets would be eroded during flight; and locations where deicing, fuel, or hydraulic fluid would come in contact with the riblets are left bare. After these concessions, the riblets covering the remaining 70 % of the aircraft have provided 3 % total drag reduction. This 3 % drag reduction correlates to a similar 3 % savings in fuel costs [4].

Another large commercial application for riblet technologies is drag reduction in pipe flow. Machining the surface or applying vinyl film riblets proves difficult in the confines of most pipes, and an alternate solution must be used. Experimental application of a scratching technique to the inside surface of pipes has created a riblet-like roughness that has provided more than 5 % drag reduction benefit [33]. Stemming from an old sailors' belief that ships sail faster when their hulls are sanded in the longitudinal direction, Weiss fabricated these riblets by using a steel brush moved through the pipeline to create a ridged surface. Studies have shown as much as a 10 % reduction in fluid flow with the combined effect of cleaning the pipe and ridging the surface. Tests on a 10-mile gas pipeline section have confirmed this benefit during commercial operation [4].

The dominant and perhaps only commercial market where riblet technology for drag reduction is commercially sold is competitive swimwear [14]. The general population became aware of shark skin's drag reduction benefits with the introduction of the FastSkin® suits by Speedo in 2004. Speedo claimed a drag reduction of several percent in a static test compared to other race suits. However, given the compromises of riblet geometry made during manufacturing, it is hard to believe the full extent of the drag reduction.

It is clear that creating surface structures by weaving threads is difficult. As a result, riblet



Shark Skin Effect, Fig. 9 Images of riblet geometries on (a) unstretched and (b) stretched Speedo FastSkin® swimsuit. (c) Schematic showing apparent hierarchical riblet structure formed by threads [14]

geometries woven from thread have limited options of feasible riblet shapes. By the pattern woven into the FastSkin® swimsuits, riblets are formed which resemble wide blade riblets with small grooves on top. The larger riblets are formed by the macro-weaving pattern, and the smaller riblets are created by the individual weaves of thread aligned with the macro-riblets. Both of these riblet-like shapes are distinguishable in Fig. 9. As shown in Fig. 9a, unstretched

riblets are tightly packed. As the fabric stretches, the riblet width and spacing increase (Fig. 9b). The associated decrease in h/s ratio depends on the dimensions of each swimmer's body, which is another compromising factor in the design. Riblet thickness is also a factor considered in the design. Aside from the limitations imposed by the weaving patterns available, flexibility in the riblet tips will hinder the fabric's ability to impede the cross-stream translation of streamwise vortices. Thicker riblets are probably needed, used for strength, and cause a decrease in the peak drag reduction capability compared to thinner riblets.

3 M vinyl film riblets [27] have been applied to many test surfaces, including the inside of various pipes for pipe flow studies [22], flat plates in flow channels and wind tunnels, boat hulls in towing tanks [11], airplane wings [17], and airplane fuselages. Similar riblet films have been fabricated using bulk micromachining of silicon to create a master for molding of Polydimethylsiloxane (PDMS) to create a thin, flexible riblet film. This film has been used in flow visualization tests [25].

Grinding and rolling methods of riblet fabrication have been studied for application in both research and large-scale application. A profiled grinding wheel has been used to fabricate several riblet geometries based on sawtooth riblets with $h = 20 \mu\text{m}$ and $s = 50 \mu\text{m}$ [15]. Dressing of the grinding wheel was done with diamond-profile roller used in a two-step process in which the profile roller dresses every second tooth on the first pass, shifts axially the distance of one riblet spacing, and dresses the remaining teeth on a second pass. One downside of the grinding process is the lack of hardening on the final riblet surface. Alternatively, rolling methods can be used to strain harden the riblets during fabrication. Using a roller with the profile of two riblets on its outer face, a linearly patterned rolling process has been used to fabricate scalloped riblets in a titanium alloy with $h = 162 \mu\text{m}$ and $s = 340 \mu\text{m}$ [21]. The strain hardening, favorable grain patterns, and residual compressive stresses in the riblet surface after fabrication provide advantages in riblet strength for production applications.

Effect of Fish Mucus and Polymers on Fluid Drag

Fish are known to secrete mucus during swimming [14]. Though it is not known whether the mucus is present at all times, it is known that certain environmental factors cause, alter, or enhance the production of mucus. These environmental stressors may present a need for increased swimming speed to catch or avoid becoming prey, for protection against non-predatory threats such as microorganisms, or to resist abrasion while swimming near rocky surfaces. Regardless of which events cause fish to secrete mucus, the drag reduction benefits during mucus-assisted swimming are known. Numerous experiments have demonstrated the drag reduction possible with fish-covered mucus compared to non-mucus-covered shapes [18]. In an experiment comparing the drag on wax models to a mucus-covered fish, a reduction in skin friction drag of 50 % was seen [13].

Similar to these fish mucus experiments, polymer additives in pipe flows have been known for many years to reduce the drag in fluid flows by extreme amounts. Polymers are known to have low shear strength which results in low friction [8, 9]. In a pipe flow study comparing various injection techniques of polymer solutions into water, drag reductions of up to 80 % were achieved [16]. Additionally, the drag reduction benefit increases with increased Reynolds number. While this works well for pipe flows, in which the polymer remains mixed and active throughout the length of the pipe, its application to external flows is much more difficult. Mucus on fish does not mix well with water in static contact, but does mix and provide drag reduction during dynamic contact. By this feature, the mucus use of fish is minimized. Unfortunately for any long range application of polymer drag reduction on an external flow, the polymer solution must be continuously injected. This would cause large quantities of the solution to be used and likely render the strategy inefficient in terms of overall energy use.

Though sharks do not secrete enough mucus to use this mechanism for drag reduction, small amounts of mucus are present on the skin of

sharks [2]. It is possible that shark skin mucus secretion is similar to fish, where only environmental stressors or swimming causes an increase in output, but the total quantity of mucus on the surface at any given time is likely quite low. One possible mechanism by which this trace quantity of mucus could be useful is in changing the flow characteristics in the riblet valleys or at the riblet peaks, where shear stresses are highest. In the riblet valleys, a trace secretion of mucus could increase flow velocity and decrease the overall momentum transfer from the shark to the surrounding water by condensing the overall structure of the velocity gradient. Alternatively, injection at the riblet peaks may cause a reduction in shear stresses where they are at their maximum and again cause a reduction in drag. These small effects near the surface may propagate into a larger benefit as the lines of constant velocity in the flow shift and condense [14].

Closure

Fluid drag in the turbulent boundary layer is in large part due to the effects of the stream-wise vortices formed in the fluid closest to the surface. Turbulence and associated momentum transfer in the outer boundary layers is in large part due to the translation, ejection, and twisting of these vortices. Additionally, the vortices also cause high velocities at the surface which create large shear stresses on the surface. Riblets impede the translation of the stream-wise vortices, which causes a reduction in vortex ejection and outer layer turbulence. In addition, riblets lift the vortices off the surface and reduce the amount of surface area exposed to the high-velocity flow. By modifying the velocity distribution, riblets facilitate a net reduction in shear stress at the surface.

Various riblet shapes have been studied for their drag-reducing capabilities, but sawtooth, scalloped, and blade riblets are most common. Drag reduction by riblet surfaces has been shown to be as high as nearly 10 % given an optimal geometry of $h/s \sim 0.5$ for blade riblets with a no-slip condition. The maximum reliable drag reduction provided by scalloped riblets and

sawtooth riblets is about 6 % at $h/s \sim 0.7$ and 5 % at ridge angle $\alpha \sim 60^\circ$, respectively. Experimentation of other shapes has provided similar benefits of around 5 % drag reduction. Though the optimum shape for drag reduction performance is blade riblets, the fragile nature of these blades makes their commercial application of little use. Scalloped and sawtooth riblets, which provide considerably less drag reduction benefit, are much stronger shapes mechanically speaking and should be used for application in environments where contact may occur with non-fluid materials.

Commercial applications of riblets include competition swimsuits, which use a thread-based riblet geometry, as well as experimental applications to airplanes. Drag reductions in riblet application have been accomplished, and flight applications have seen fuel savings of as much as 3 %. Manufacturing techniques for riblets must also be chosen specific to their application. Vinyl film riblets are the easiest method, as application of a film to a surface requires less for work small-scale application than other methods. Rolling or grinding methods of riblet application should be investigated for turbine blades or high volume commercially sold pieces.

Cross-References

- [Biognosis](#)
- [Biomimetics](#)
- [Biomimicry](#)

References

1. Becher, D.W., Hoppe, G.: On the drag reduction of the shark skin. Paper # AIAA-85-0564, presented at AIAA Shear Flow Control Conference, Boulder, 12–14 Mar (1985)
2. Bechert, D.W., Bartenwerfe, R.M., Hoppe, G., Reif, W.-E.: Drag reduction mechanisms derived from shark skin. Paper # ICAS-86-1.8.3, Proceedings of the 15th ICAS congress, vol. 2 (A86-48-97624-01), pp. 1044–1068. AIAA, New York (1986)
3. Bechert, D.W., Hoppe, G., van der Hoven, J.G.T., Makris, R.: The Berlin oil channel for drag reduction research. *Exp. Fluids* **12**, 251–260 (1992)

4. Bechert, D.W., Bruse, M., Hage, W., Meyer R.: Biological surfaces and their technological application – laboratory and flight experiments on drag reduction and separation control. Paper # AIAA-1997-1960, presented at AIAA 28th Fluid dynamics conference, Snowmass Village, 29 June–2 Aug (1997)
5. Bechert, D.W., Bruse, M., Hage, W., van der Hoeven, J.G.T., Hoppe, G.: Experiments on drag reducing surfaces and their optimization with an adjustable geometry. *J. Fluid Mech.* **338**, 59–87 (1997)
6. Bechert, D.W., Bruse, M., Hage, W.: Experiments with three-dimensional riblets as an idealized model of shark skin. *Exp. Fluids* **28**, 403–412 (2000)
7. Bechert, D.W., Bruse, M., Hage, W., Meyer, R.: Fluid mechanics of biological surfaces and their technological application. *Naturwissenschaften* **87**, 157–171 (2000)
8. Bhushan, B.: Principles and Applications of Tribology. Wiley, New York (1999)
9. Bhushan, B.: Introduction to Tribology. Wiley, New York (2002)
10. Bhushan, B.: Biomimetics: lessons from nature – an overview. *Phil. Trans. R. Soc. A* **367**, 1445–1486 (2009)
11. Choi, K.S., Gadd, G.E., Pearcey, H.H., Savill, A.M., Svensson, S.: Tests of drag-reducing polymer coated on a riblet surface. *Appl. Sci. Res.* **46**, 209–216 (1989)
12. Choi, K.S., Yang, X., Clayton, B.R., Glover, E.J., Altar, M., Semenov, B.N., Kulik, V.M.: Turbulent drag reduction using compliant surfaces. *Proc. R. Soc A* **453**, 2229–2240 (1997)
13. Daniel, T.L.: Fish mucus: *in situ* measurements of polymer drag reduction. *Biol. Bull.* **160**, 376–382 (1981)
14. Dean, B., Bhushan, B.: Shark-skin surfaces for fluid-drag reduction in turbulent flow: a review. *Phil. Trans. R. Soc. A* **368**, 4775–4806 (2010); 5737
15. Denkena, B., de Leon, L., Wang, B.: Grinding of microstructures functional surfaces: a novel strategy for dressing of micropatterns. *Prod. Eng.* **3**, 41–48 (2009)
16. Frings, B.: Heterogeneous drag reduction in turbulent pipe flows using various injection techniques. *Rheol. Acta* **27**, 92–110 (1988)
17. Han, M., Huh, J.K., Lee, S.S., Lee, S.: Micro-riblet film for drag reduction. In: Proceedings of the Pacific Rim Workshop on transducers and micro/nano technologies, Xiamen (2002)
18. Hoyt, J.W.: Hydrodynamic drag reduction due to fish slimes. *Swim. Fly. Nat.* **2**, 653–672 (1975)
19. Jung, Y.C., Bhushan, B.: Biomimetic structures for fluid drag reduction in laminar and turbulent flows. *J. Phys.: Condens. Matter* **22**, #035104 (2010)
20. Kline, S.J., Reynolds, W.C., Schraub, F.A., Runstadler, P.W.: The structure of turbulent boundary layers. *J. Fluid Mech.* **30**, 741–773 (1967)
21. Klocke, F., Feldhaus, B., Mader, S.: Development of an incremental rolling process for the production of defined riblet surface structures. *Prod. Eng.* **1**, 233–237 (2007)
22. Koury, E., Virk, P.S.: Drag reduction by polymer solutions in a riblet-lined pipe. *Appl. Sci. Res.* **54**, 323–347 (1995)
23. Krieger, K.: Do pool sharks really swim faster? *Science* **305**, 636–637 (2004)
24. Lang, A.W., Motta, P., Hidalgo, P., Westcott, M.: Brisstled shark skin: a microgeometry for boundary layer control? *Bioinspir. Biomim.* **3**, 1–9 (2008)
25. Lee, S.J., Choi, Y.S.: Decrement of spanwise vortices by a drag-reducing riblet surface. *J. Turbul.* **9**, 1–15 (2008)
26. Lee, S.-J., Lee, S.-H.: Flow field analysis of a turbulent boundary layer over a riblet surface. *Exp. Fluids* **30**, 153–166 (2001)
27. Marentic, F.J., Morris, T.L.: Drag reduction article. US Patent 5,133,516, 1992
28. Shephard, K.L.: Functions for fish mucus. *Rev. Fish Biol. Fish.* **4**, 401–429 (1994)
29. Walsh, M.J.: Drag characteristics of v-groove and transverse curvature riblets. *Viscous Flow Drag. Red.* **72**, 169–184 (1980)
30. Walsh, M.J.: Turbulent boundary layer drag reduction using riblets. Paper # AIAA-82-0169, presented at AIAA 20th Aerospace sciences meeting, Orlando, 11–14 Jan 1982
31. Walsh, M.J., Anders, J.B.: Riblet/LEBU research at NASA Langley. *Appl. Sci. Res.* **46**, 255–262 (1989)
32. Walsh, M.J., Lindemann, A.M.: Optimization and application of riblets for turbulent drag reduction. Paper # AIAA-84-0347, presented at AIAA 22nd aerospace sciences meeting, Reno, 9–12 Jan 1984
33. Weiss, M.H.: Implementation of drag reduction techniques in natural gas pipelines. Presented at 10th European drag reduction working meeting, Berlin, 19–21 Mar 1997
34. Wilkinson, S.P.: Influence of wall permeability on turbulent boundary-layer properties, Paper # AIAA 83–0294, presented at 21st aerospace sciences meeting of the american institute of aeronautics and astrodynamics, Reno, 10–13 Jan 1983
35. Wilkinson, S.P., Lazos, B.S.: Direct drag and hot-wire measurements on thin-element riblet arrays. Presented at the IUTAM symposium on turbulence management and relaminarization, Bangalore, 19–23 Jan 1987
36. Wilkinson, S.P., Anders, J.B., Lazos, B.S., Bushnell, D.M.: Turbulent drag reduction research at NASA Langley: progress and plans. *Inter. J. Heat Fluid Flow* **9**, 266–277 (1988)

Shark Skin Separation Control

► [Shark Skin Drag Reduction](#)

Short (Low Aspect Ratio) Gold Nanowires

► Gold Nanorods

Short-Interfering RNA (siRNA)

► RNAi in Biomedicine and Drug Delivery

Shrinkable and Stretchable Nanomanufacturing

Clifford J. Engel and Teri W. Odom
Department of Chemistry, Northwestern University, Evanston, IL, USA

Definition

Shrinkable and stretchable nanomanufacturing extends the capabilities of soft lithography by allowing dynamic control of nanopatterns without the need to remake a master.

Overview

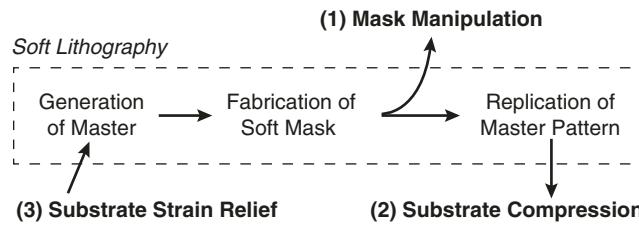
Early advances in nanomanufacturing were driven by innovations in the electronics industry. The invention of the integrated circuit in 1958 by Jack Kilby at Texas Instruments created a need for manufacturing tools that could pattern multiple components (e.g., transistors, interconnects, and other functional elements) on a single substrate. While the first commercially available chips did not have nano-sized components, future progress required the ability to reduce the size of components. For example, the deep-UV excimer laser lithography system was introduced in 1982, which overcame the 500 nm resolution limits of mercury-xenon lamps while increasing throughput by two orders of magnitude [1]. Advances in shorter-wavelength light sources were critical in

improving the resolution of photolithography from 500 nm in 1990 to 45 nm and below in 2000 [2].

For emerging industries and consumer products besides chip-based electronics, like DNA microarrays and wearable sensors, a primary goal of nanomanufacturing has not been simply miniaturization but the ability to pattern nonplanar and soft material substrates at reduced costs. Thus, alternative strategies, such as soft lithography, to conventional microfabrication methods were developed [3]. Soft lithography offers a conceptually different approach to nanofabrication and consists of two major steps: (1) the fabrication of a master and (2) the replication of the master pattern. Typically, a master is generated using tools from the microelectronics industry, and the replication process involves a soft, elastomeric mask created by molding a polymer against a master pattern. The most common mask material, poly(dimethylsiloxane) (PDMS), is commercially available as Sylgard 184 (Dow Corning) and can replicate feature sizes that are 500 nm and larger with high fidelity. Composite stamps consisting of a thin, hard PDMS layer supported on a flexible PDMS backing can extend the pattern transfer and replication capabilities down to 50 nm in lateral dimensions [3] and < 2 nm in the vertical dimension [4]. To create replicas of the master on a desired surface, the soft mask is brought into contact with a patterning material. As a result, the PDMS mask determines the structural and geometrical characteristics of the molded and printed material.

The patterning material can vary depending on the desired chemical or physical properties of the pattern for an application. Printing techniques such as microcontact printing transfer the 2D structure of the PDMS mask by bringing molecular “ink” into contact with a metal thin film. Monolayers of alkanethiols can be printed on noble and coinage metals (Au, Ag, Pd, and Pt), while alkylsiloxanes can be printed on Si or glass. The end functional groups of the molecular inks control chemical properties of the patterned surface, such as hydrophobicity or reactivity [3].

In contrast to printing, molding retains the 3D structure of the mask, where the precursors are



Shrinkable and Stretchable Nanomanufacturing,

Fig. 1 Overview of shrinkable and stretchable nanomanufacturing: Soft lithography processes involve generation of a master pattern through photolithography, fabrication of a soft mask molded against the master

usually liquid prepolymers such as UV-curable polyurethane or thermally curable epoxies [3]. The precursor material undergoes a transition from liquid to solid while in direct physical contact with the PDMS mask. The prepolymer is selected for optimal physical properties such as optical transparency or elasticity. Despite the versatility of soft lithography, the end goal of this suite of techniques is to replicate or duplicate the pattern on the master. Applications that require a range of structures, therefore, need to make a new master for each desired pattern, which can be time-consuming, tedious, and costly.

A new class of techniques—shrinkable and stretchable nanomanufacturing—can expand the capabilities of soft lithography to generate patterns without needing to remake the master or, in some cases, eliminate the need for a master entirely (Fig. 1). This article will describe how such patterns can be achieved by (1) manipulating the mask during the replication process, (2) post-processing the patterned substrate by compressing or stretching, and (3) avoiding a mask and relying on the spontaneous generation of patterns from unpatterned films.

Manipulating the Soft Mask

Although PDMS is a ubiquitous elastomer for preparing masks in soft lithography, a major disadvantage is that it swells when in contact with nonpolar solvents [5]. Swelling of PDMS distorts the mask and is a major limitation for techniques that directly replicate the mask pattern. The

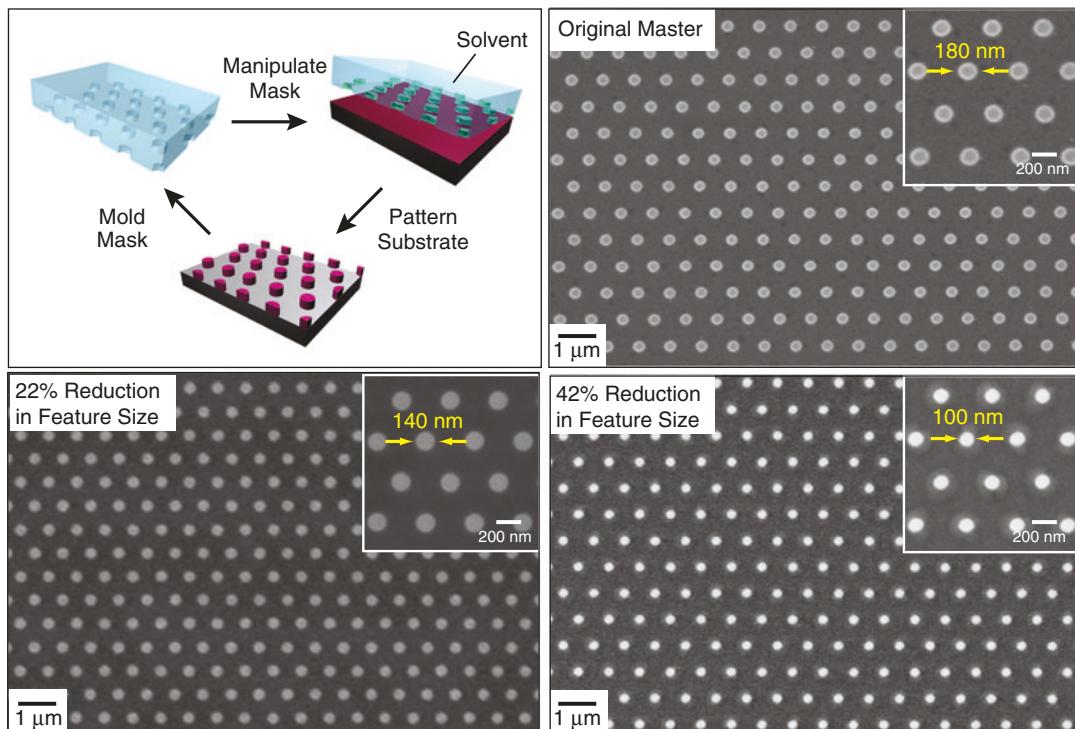
pattern, and replication of the master pattern into the patterning substrate. Modifying the nanopattern can be accomplished by (1) manipulating the soft mask, (2) compressing the substrate after patterning, or (3) relieving the strain in a thin film to form spontaneous features

swelling ratio S (the ratio of the length of PDMS in solvent over the length of dry PDMS) can be as large as 2.13 in diisopropylamine, 1.30 in toluene, and 1.22 in chlorobenzene [5]. As a result, most procedures use only non-swelling solvents such as water, dimethylformamide, or ethanol ($S = 1.00\text{--}1.02$) [5].

Solvent-assisted nanoscale embossing (SANE) is an all-moldable nanofabrication platform that has exploited the swelling properties of PDMS to produce, from a single master, nanoscale patterns with programmable feature sizes. Using solvents with specific S , the degree of swelling can shrink the size of the features accordingly in the mask during the embossing process *without* affecting the periodicity of the patterns (Fig. 2). For example, changing the solvent from a low-swelling solvent like dimethylformamide to a high-swelling solvent like dichloromethane can reduce feature sizes from 140 nm to 100 nm ($\approx 30\%$). Since swelling is a reversible process, a single PDMS mask can be dried and soaked in different solvents during iterative embossing processes to generate patterned surfaces with varying feature sizes without different master patterns [6].

Post-processing of Patterned Substrates

If patterns are created on hard or stiff substrates such as Si, the separation between features cannot be modified post-patterning. In contrast, patterning on soft substrates enables possibilities to tune the spacing and symmetry of features by compressing or stretching the substrate. Although



Shrinkable and Stretchable Nanomanufacturing,
Fig. 2 Manipulating the soft mask: Solvent-assisted nanoscale embossing (SANE) involves the swelling of a PDMS mask with varying solvents during the replication process. The swelling of PDMS reduces the feature sizes of the patterned substrate without changing the periodicity of the master pattern. The patterned substrate with reduced

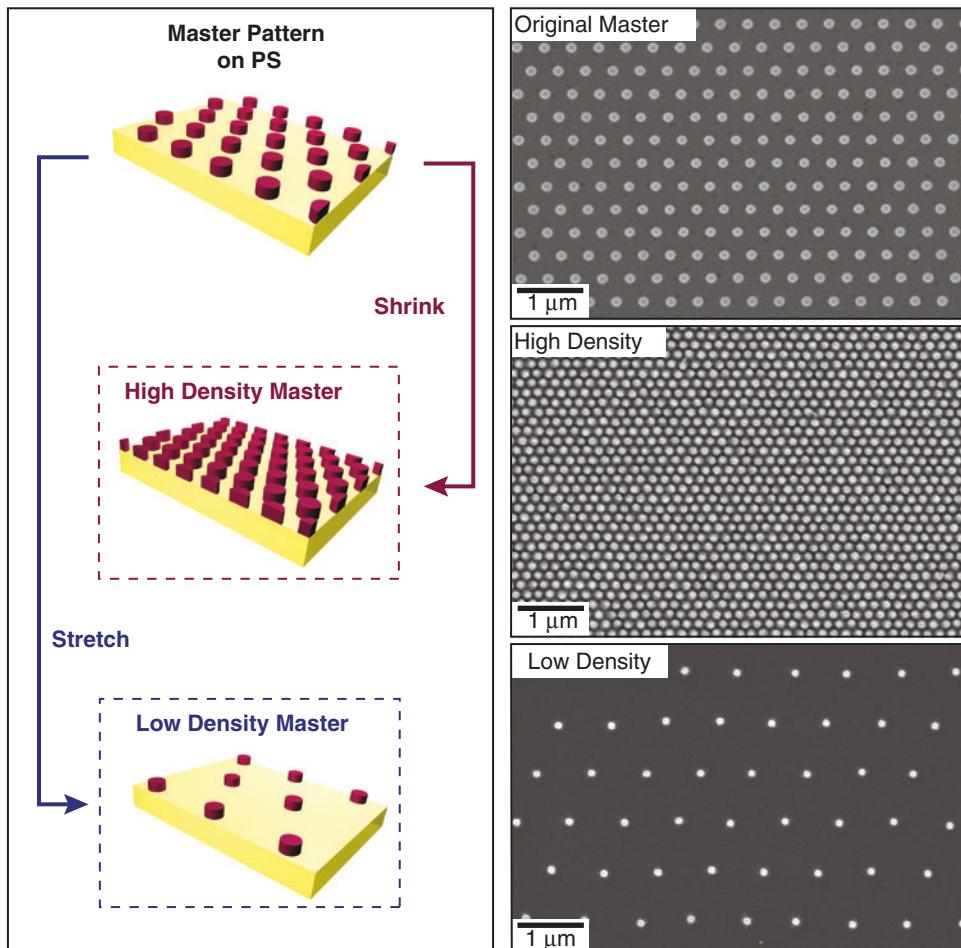
feature size can act as a template for molding a new mask and repeating the process. Feature sizes were reduced through the replacement of the non-swelling solvent ethanol (*original master*) with high-swelling dimethylformamide (22 % reduction) and dichloromethane (42 % reduction) solvents

PDMS is one option for such substrates [7], the elastomer PDMS returns back to its original dimensions when the strain is removed. In contrast, thermoplastics such as polystyrene (PS) require no external force to maintain the stability of stretched or compressed substrates once heated and cooled above their glass transition temperature (T_g), representing an advantage over PDMS.

A stretchable and shrinkable manufacturing technique—inverse solvent-assisted nanoscale embossing (inSANE)—effectively SANE on thermoplastics, can maintain the patterned feature sizes dictated by the PDMS mask yet with tunable spacing and symmetry [6]. Features patterned with inSANE on a PS surface can undergo controlled shrinking to reduce feature periodicity by up to 50 % (Fig. 3). Importantly, this pattern with

reduced spacing can now function as a new master for subsequent use in soft lithography. The major breakthrough of SANE/inSANE is that a single master pattern can now be used to create a near-infinite range of master molds with varying feature size, spacing, and orientation. No other existing nanopatterning methods can both prototype arbitrary patterns with small separation and scale for less than \$100.

In addition, shrinkable and stretchable nanomanufacturing can also be incorporated into other nanofabrication techniques, such as patterning metals via deposition through a shadow mask. The size of the patterned metal is determined by the size of the gaps in the mask (limited to ca. 100 nm by photolithography). Replacing the mask material with a thermoplastic enables reduction of the gap size through controlled shrinking.



Shrinkable and Stretchable Nanomanufacturing, Fig. 3 Post-processing of patterned substrates: The master pattern is first replicated on a thermoplastic PS. After heating above T_g , the thermoplastic can be either shrunk to increase the pattern density or stretched to decrease the

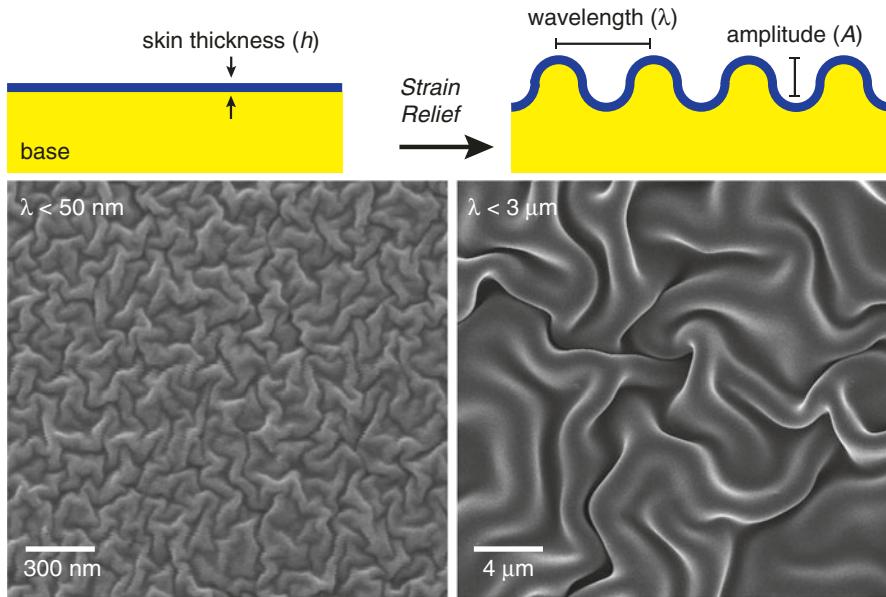
pattern density. In a single step, the master pattern periodicity (400 nm) is reduced to 200 nm (*high density*) by shrinking or increased to 800 nm (*low density*) by stretching

For example, the final gap size depends on the time and temperature at which the thermoplastic was heated above T_g and not the feature size of the initial slit patterning controlled by photolithography. As a result, controlling the shrinking of the mask enables the tuning of patterned metal features from 2 μm down to 21 nm [8].

Nanopatterns Without Needing a Mask

When a thin film (skin) bound to a soft layer (base) is compressed, out-of-plane buckles can

form. The formation of such features is known as wrinkling, a spontaneous and mask-less process whose patterned area is scalable to hundreds of square meters. For comparison, as of 2015, the largest photomask available for purchase for fabricating PDMS masters is 32" by 32" with feature resolution down to 1 μm [9]. Useful as a mask-less method, the out-of-plane buckles are periodic, where the distance between features is characterized by an average wrinkle wavelength (λ). The mechanical properties of the skin and base prior compression dictate the final wrinkle λ . Changing the thickness of the skin and



Shrinkable and Stretchable Nanomanufacturing,
Fig. 4 *Nanopatterns without needing a mask:* After applying strain relief to a stiff skin (h) on a soft base, periodic out-of-plane buckling will occur. The wrinkles are characterized by their wavelength (λ) and amplitude

(*4*). The wrinkle wavelength of a polymer stiff skin grown on a thermoplastic substrate can be continuously tuned from 30 nm to several μm by changing the skin thickness prior to strain relief

skin/base materials can tune the wavelength from tens of nanometers to hundreds of meters. Wrinkles have been incorporated into applications like organic photovoltaics [9], flexible electronics [10], and fluorescence sensing of biomolecules [11], which benefit from large-area patterns.

The selection of materials for both the skin and base controls the geometric characteristics of the wrinkle patterns. For example, the deposition of thin metal (Au) films on a thermally expanded PDMS base can result in wrinkles with λ from 100 μm down to 4 μm upon the PDMS relaxing back to its normal size at room temperature. The thickness of the Au controls the final wrinkle λ . Unfortunately, during the physical deposition of Au onto the PDMS base, the top region stiffens, which ultimately limits the lower limit of skin thickness and resultant wavelengths of this system [12]. To overcome this limitation, the metal skin can be replaced with a silica-like film formed on PDMS by plasma-mediated growth. By using an

oxygen plasma to control silica skin thickness, submicron wrinkles ($\lambda \approx 500 \text{ nm}$) were formed [12]. Ultraviolet/ozone (UVO) radiation (skin $\approx 5 \text{ nm}$) of the PDMS slab has resulted in a modified PDMS wrinkle systems with λ as low as 80 nm [12].

Alternative wrinkle systems have replaced the PDMS base layer with thermoplastics such as PS in order to expand the capabilities of wrinkle fabrication. For example, a Au film deposited on PS, when heated above T_g , can generate wrinkles with λ from several microns down to 250 nm, four times smaller than for Au of a similar thickness on PDMS [13]. These lower wavelengths, which approach the nanoscale regime, are possible as the Young's moduli ratio of skin and base (E_S/E_B) decreases. When the Au film is replaced with a polymer skin grown on a thermoplastic substrate, for example, wrinkle λ can be continuously tuned down to a lower limit of $\lambda \approx 30 \text{ nm}$ (Fig. 4). The use of a polymer thin film for the stiff skin resulted in a 6 times reduction in E_S/E_B compared

to silica on PDMS [14]. Pushing the lower limits of nanowrinkle fabrication can be achieved by optimization of the mechanical properties of the skin and base.

Besides λ , the amplitude A plays an important role in determining the physical properties of a nanowrinkled substrate. Unlike other nanopatterns, wrinkles have a sinusoidal nature, where A is reduced when the vertical feature size is reduced. A can be maximized for a given wavelength by increasing the applied strain until a critical strain threshold is reached, after which nonlinear effects cause structural breakdowns and folds, tears, or self-similar wrinkles form [12]. The critical strain threshold limits A in most wrinkle systems, including PDMS, to ca. 0.1 λ . As a result, nanowrinkles ($\lambda \leq 100$ nm) are limited to A less than 10 nm. Alternative wrinkle material systems can overcome this limitation. For example, the polymer skin grown on thermoplastic PS has advantages since both the polymer skin and base can mitigate nonlinear effects at high strains, increasing the critical strain threshold to over 50 %. As a result, wrinkles with $A > \lambda$ can be formed, which is a tenfold increase in amplitude over PDMS wrinkle systems with similar λ [15].

Hybrid Mask and Mask-Free Techniques

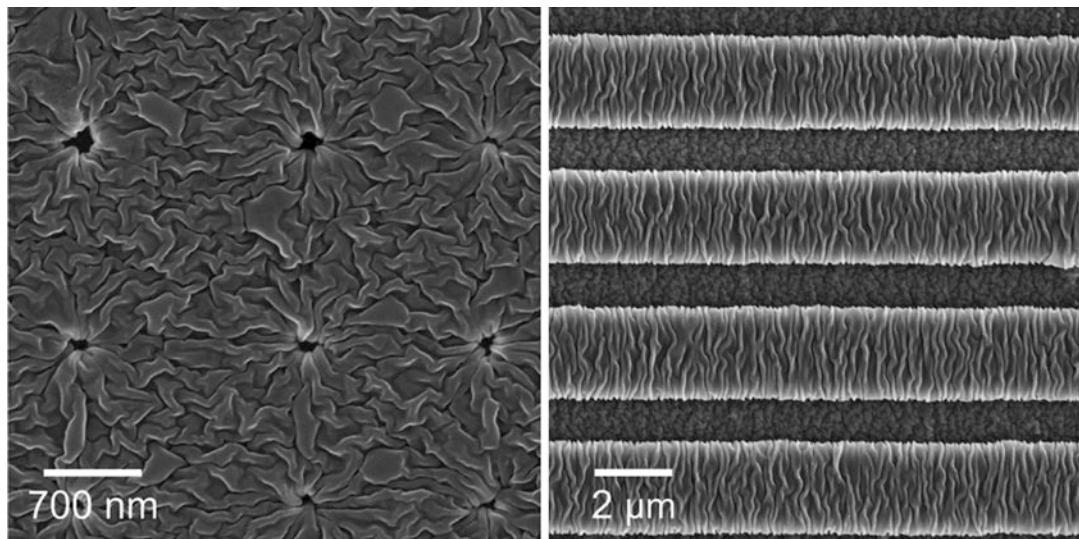
Hybrid nanofabrication approaches combine two or more techniques to produce nanopatterns not feasible with a single method and, importantly, can also overcome limitations of the individual techniques. For example, a drawback of wrinkles is that they are disordered within a characteristic wavelength; however, patterning the stiff skin prior to shrinking will orient the wrinkles perpendicular to a patterned feature [12]. This controlled order is a result of nonuniform local strain at the patterned/unpatterned interface; wrinkles will remain ordered for distances 20 λ –100 λ away from the patterned feature [12, 15]. Disrupting the local strain to introduce order will have a negative effect on the wrinkle A and structure. As a result, the ratio between patterned feature

widths (w) and wrinkle λ must be controlled to optimize the final wrinkle structure. At the micron scale, patterning lines with 200 μm to 500 μm w into a Au film prior to applying compressive force produced ordered wrinkles with wavelengths of 20–50 μm [12]. To order sub-100 nm wrinkles, inSANE was combined with wrinkling on PS. The inSANE pattern acted as a mask for the growth of the patterned polymer skin while the thickness of the polymer skin determined the final wrinkle λ , resulting in ordered 50 nm wrinkles [15]. The A and wrinkle order was measured with varying line patterns and skin thicknesses to optimize the relationship between λ and w . It was found that as w approached 20 λ , wrinkles would become ordered without any observable decreases in A . As w was further decreased to 2 λ , the strain manipulation became large enough to collapse the wrinkles into folds (Fig. 5). Since nanowrinkles and nanofolds spontaneously form across the entire surface, a hybrid scheme provides a parallel method to create complex ordered and disordered patterns with nanoscale spatial resolution.

Hybrid nanofabrication techniques can also be used to form hierarchical patterns. Multi-scale structures are abundant in nature; macroscale features impart mechanical stability, and nanoscale features control functionality such as wettability, optical properties, and adhesion [13]. Control over features at all length scales is key to fabricating engineered systems that can mimic or eventually outperform natural systems. Physical imprinting followed by swelling of a soft polymer produced a hierarchical pattern with nanoscale lines (500 nm) and macroscale wrinkle features ($\lambda = 15$ –35 μm) [16]. In addition to forming a hierarchical structure, controlling the embossing depth and direction of the nanoscale lines tuned the previously uncontrollable periodicity and orientation of the macroscale wrinkles.

Future of Shrinkable and Stretchable Nanomanufacturing

Shrinkable and stretchable strategies enable unique applications such as the manipulation of



Shrinkable and Stretchable Nanomanufacturing,
Fig. 5 Hybrid mask and mask-free techniques: Combination of inSANE and nanowrinkle fabrication results in ordered nanowrinkles. The shape and periodicity of the nanoembossed pattern determine the degree of wrinkle

order. With a single patterned feature, the wrinkles orient radially around the hole. When the hole is replaced with a line with a width less than 20λ , wrinkles become completely ordered

cell adhesion with nanopatterns [17], the formation of large-area antibacterial surfaces [18], and the fabrication of bioinspired hydrophobic surfaces [13]. Designs for these emerging fields require patterning techniques that can control feature shapes and sizes from 10^{-9} to 10^{-3} m while simultaneously reducing costs associated with nanopatterning [13], a major strength of shrinkable and stretchable nanomanufacturing. The development of simple yet transformative approaches that can achieve independent control over feature size, periodicity, and orientation of patterns is the future of nanomanufacturing.

Cross References

- [Flexible Electronics](#)
- [Mechanical Properties of Nanocrystalline Metals](#)
- [Micro- and Nanomanipulation for Nanomanufacturing](#)
- [Microcontact Printing](#)
- [Nanoimprinting](#)

- [Nanomechanical Properties of Nanostructures](#)
- [Nanotechnology](#)
- [Reliability of Nanostructures](#)

References

1. Xia, Y., Rogers, J.A., Paul, K.E., Whitesides, G.M.: Unconventional methods for fabricating and patterning nanostructures. *Chem. Rev.* **99**, 1823–1848 (1999)
2. Basting, D., Marowsky, G.: *Excimer laser technology*. Springer Science & Business Media, Berlin (2005)
3. Qin, D., Xia, Y., Whitesides, G.M.: Soft lithography for micro- and nanoscale patterning. *Nat. Protoc.* **5**, 491–502 (2010)
4. Gates, B.D., Whitesides, G.M.: Replication of vertical features smaller than 2 nm by soft lithography. *J. Am. Chem. Soc.* **125**, 14986–14987 (2003)
5. Lee, J.N., Park, C., Whitesides, G.M.: Solvent compatibility of poly(dimethylsiloxane)-based microfluidic devices. *Anal. Chem.* **75**, 6544–6554 (2003)
6. Lee, M.H., Huntington, M.D., Zhou, W., Yang, J.-C., Odom, T.W.: Programmable soft lithography: solvent-assisted nanoscale embossing. *Nano Lett.* **11**, 311–315 (2011)
7. Xia, Y., Whitesides, G.M.: Reduction in the size of features of patterned SAMs generated by microcontact printing with mechanical compression of the stamp. *Adv. Mater.* **7**, 471–473 (1995)

8. Zhang, B., Zhang, M., Cui, T.: Low-cost shrink lithography with sub-22 nm resolution. *Appl. Phys. Lett.* **100**, 133113 (2012)
9. Kim, J.B., Kim, P., Pégard, N.C., Oh, S.J., Kagan, C. R., Fleischer, J.W., Stone, H.A., Loo, Y.-L.: Wrinkles and deep folds as photonic structures in photovoltaics. *Nat. Photon* **6**, 327–332 (2012)
10. Rogers, J.A., Someya, T., Huang, Y.: Materials and mechanics for stretchable electronics. *Science* **327**, 1603–1607 (2010)
11. Sharma, H.; Digman, M. A.; Felsinger, N.; Gratton, E.; Khine, M.: Enhanced emission of fluorophores on shrink-induced wrinkled composite structures. *Opt Mater Express* **4**, 753–763 (2014)
12. Genzer, J., Groenewold, J.: Soft matter with hard skin: from skin wrinkles to templating and material characterization. In *Soft Matter* **2**, 310 (2006)
13. Bae, W.-G., Kim, H.N., Kim, D., Park, S.-H., Jeong, H.E., Suh, K.-Y.: 25th Anniversary article: scalable multiscale patterned structures inspired by nature: the role of hierarchy. *Adv. Mater.* **26**, 675–700 (2013)
14. Huntington, M.D., Engel, C.J., Hryny, A.J., Odom, T. W.: Polymer nanowrinkles with continuously tunable wavelengths. *ACS Appl. Mater. Interfaces* **5**, 6438–6442 (2013)
15. Huntington, M.D., Engel, C.J., Odom, T.W.: Controlling the orientation of nanowrinkles and nanofolds by patterning strain in a thin skin layer on a polymer substrate. *Angew. Chem. Int. Ed.* **53**, 8117–8121 (2014)
16. Li, Y.; Dai, S.; John, J.; Carter, K. R.: Superhydrophobic surfaces from hierarchically structured wrinkled polymers. *ACS Appl. Mater. Interfaces* **5**, 11066–11073 (2013)
17. Yang, P., Baker, R.M., Henderson, J.H., Mather, P.T.: In vitro wrinkle formation via shape memory dynamically aligns adherent cells. *Soft Matter* **9**, 4705–4714 (2013)
18. Freschauf, L.R., McLane, J., Sharma, H., Khine, M.: Shrink-induced superhydrophobic and antibacterial surfaces in consumer plastics. *PLoS One* **7**, e40987 (2012)

Si Nanotubes

► [Physical Vapor Deposition](#)

Silent Flight of Owls

► [Silent Owl Wings](#)

Silent Owl Wings

Thomas Bachmann¹ and Hermann Wagner²

¹Institute for Fluid Mechanics and Aerodynamics, Technische Universität Darmstadt, Darmstadt, Germany

²Institute for Biology II, RWTH Aachen University, Aachen, Germany

Synonyms

[Silent flight of owls](#)

Definition

Wings of owls are equipped with macro- and microstructured devices that reduce flight noises significantly. These birds of prey hunt in both twilight and darkness. Since visual information is limited at that time of day, owls use acoustic information to detect and track potential prey. The silent flight affords the detection of prey in flight and makes the owl inaudible for its prey. Different wing and feather specializations, such as serrations at the leading edge of the wing, a velvet-like upper surface and fringes along the inner vanes, influence the airflow and thus are important for the reduced flight noise in owls.

Overview

S

Fundamentals of Bird Flight

In the course of evolution, birds have conquered the air for locomotion. A great variety of wing geometries have evolved for different needs of birds. Wings can be long or short, narrow or broad, thick or thin. Furthermore, wings can be pointed or have a rounded tip, which may be smooth or fingered. Considering this variety, each wing appears to be adapted to the distinct flight conditions of the particular species and the ecological niche it uses. However, all wings have at least one aerodynamic feature in common: They are cambered to the dorsal side. Only curved

wings are able to produce enough lift at moderate angles of attack and low flight speeds to allow a bird to become airborne. Bird wings are mostly cambered in proximal regions. Thickness as well as camber decrease towards the wing tip. The resulting reduction of mass at the wing tip causes a decrease of inertia to allow high wing-beat frequencies (2–12 Hz) [1] at low energy consumption [2].

A bird has to defy gravity to become airborne. Hence, flight consumes a lot of energy, especially during takeoff and landing [2]. High lift is achieved either by an intense beating of the wings or by forming wings with high-lift properties [3]. Such wings are characterized by a high camber and a large wing area. Apart from gravity, birds have to deal with an additional force, the drag. Drag is the resistance of the body and wings while moving through the air. It is comprised of three components [3]: first, the friction drag between the airflow and the surface, second, the form drag of the bird's body and the leading edges of the wings, and third, the induced drag. The induced drag is the component of drag force on the bird's wing that is caused by an induced downward velocity. The drag is also influenced by the microstructure of the plumage. Several anatomical specializations of feathers have been reported that influence the airflow around the bird wing, most of them in owls [4–8].

Function of Silent Flight in Owls

Owls have an almost global distribution, which is reflected in the variety of species and subspecies. These birds spend a high proportion of their time searching for prey, either by sitting on a perch, or flying slowly over fields and vegetation [9, 10]. The prey of most owls is mainly active at night. Since visual information is limited in low-light conditions, owls use acoustical cues to detect their prey. The prey emits noise by rustling through the vegetation or through intra- and interspecific communication sounds which can be detected by the owl even in absolute darkness [11]. Several anatomical and behavioral adaptations of the hunting strategy have been discovered in owls. One conspicuous specialization in many owl species is the facial ruff with its specialized

feathers that direct sound toward the ears like a dish antenna. The facial ruff is also responsible for the low hearing threshold [12]. Asymmetrically arranged ear flaps in front of the outer ear canals help to localize the elevation of sound sources. Furthermore, nuclei in the brain that process acoustic information are enlarged compared with other similar-sized birds [13]. With this set of adaptations, owls are able to detect and localize their prey precisely. Nevertheless, to be able to utilize this ability during flight, owls should fly silently. If that would not be the case, noise emitted by prey would be masked by the owl's own flight noise and the prey might be warned by the approaching owl. Consequently, owls rely on a silent flight, which is achieved by adaptations of the wing, the plumage, and the flight behavior [4–10].

Specializations of the Owls' Plumage

Feathers are the main aerodynamic components of a bird's wing. Interaction with the airflow during flight and protection against cold, heat, and wetness are essential functions. Furthermore, feathers are used for display or camouflage. A feather consists of a central shaft (rachis) and two laterally attached vanes. The feather vane is made up of parallel barbs that branch off from either side of the rachis. Barbs are interlinked via hook and bow radiates to form a closed surface. While each feather quill is inserted into the integument, all flight feathers (remiges) are additionally associated with the underlying skeleton. Secondary remiges are connected to the posterior edge of the ulna; primary remiges are supported by bones of the hand and fingers. Coverts arise from the anterior integument membrane, forming smooth and closed upper and lower wing surfaces. All wing feathers together form a wing that is light and flexible, but strong enough to meet the bird's requirements to fly.

In owls, several microstructured plumage specializations are known that influence the flow over the wing [4–8]: The dorsal surface of the feathers has an increased roughness caused by elongations of hook radiates, termed pennula. Those feathers that form the leading edge of the distal wing are equipped with serrations at their outer vanes.

Every feather of the wing is surrounded by a fringed structure which is due to unconnected barb endings. Furthermore, owl feathers are very flexible and pliant due to a reduced number of radiates and thus a lower number of interlinks between the barbs compared to other birds [8]. Finally, cross-sectional profiles of the rachises are less structured compared to other birds of similar size. These plumage specifications have been implicated in noise suppression during flight by damping noise above 2 kHz [5]. Thus, flight noise is reduced within the typical hearing spectrum of the owls' prey (>3 kHz) [5] and within the owl's own best hearing range (5–9 kHz) [14]. The reduction of noise is mainly achieved by airflow control and friction reduction of single feathers.

Basic Methodology

Morphometrics

Bird wings, together with many other parameters, have been morphometrically characterized in order to classify bird species and to create a systematic order within birds (*Aves*). However, aerodynamic parameters are not included in this approach. Aerodynamic properties such as wing span, wing area, and aspect ratio can be measured in living birds or well-preserved museum specimens. Camber and thickness distribution are more prone to errors due to drying processes or unnatural fixation. Therefore, anesthetized or recently deceased birds yield better results. Best results are gained from experiments with living birds. Hereby, the wing geometry is obtained during free-flight experiments. Computer-aided analyses of three-dimensional reconstructions of the wings provide high-resolution profile data.

Free-flight experiments are conducted to measure the noise emission of flying owls in different flight modes while the birds fly over an extreme sensible microphone array at a defined flight speed and height. The measured noise spectrum can then be correlated with the flight parameters. Finally, the results are compared to other non-silently flying birds [15].

Substructures of the feathers are investigated with different digital imaging techniques having

high spatial resolutions. Light microscopy and scanning electron microscopy (SEM) enable to some extent two-dimensional measurements and morphometric characterizations. Confocal laser scanning microscopy (CLSM) and micro-computed-tomography (μ -CT) allow the precise digitization of the three-dimensional shape of microstructures. In all cases, the data are then processed and analyzed with adequate visualization and measuring software to determine relevant morphometric and aerodynamic parameters. After specifying the shape and geometry of the owl-specific structures, these can be artificially manufactured and applied to wing models for further analysis, e.g., in wind tunnel experiments. By this means, their aerodynamic function may be clarified.

Behavioral Studies

Birds use their wings to maneuver through the air. The form and geometry but also the movements of the wings are important for an efficient locomotion. Hence, studying the flight behavior of birds is one fundamental piece of the large puzzle of understanding how bird wings work. The study of wing motion in different flight modes, ideally in combination with flow visualization, helps to understand the functions of each wing component. Flight speed, wing-beat frequency, and wing-beat amplitude are some parameters that influence the flight of birds. Different flight modes like flapping flight, gliding, or soaring exist. Furthermore, some birds adapt their flight behavior during hunting, escaping, courtship, and other modes. Nowadays, high-speed cameras outperform traditional bird watching with binoculars and cameras in many aspects. Images with high temporal and spatial resolution of wing and feather movements during beating of the wings allow a much better investigation of each wing element and its aerodynamic function. Such techniques are used in field observations or wind tunnel experiments with living birds.

Wind Tunnel Experiments

Fixed wings of birds and artificial models of bird wings are investigated in wind tunnel

experiments. Performing experiments in wind tunnels guarantee reliable and repeatable results. The use of wing models instead of living animals allows the manipulation of certain parameters of the wing. The impact of several wing geometries can then be compared and general conclusions can be drawn based on these results. Flow visualization techniques such as particle-image velocimetry (PIV), oil-flow pattern or pressure measurements are carried out to understand the influence of the wing profiling and each wing element on the airflow. In case of owl wings, the specific feather adaptations are added separately on a wing model or are removed step by step from fixed owl wings. Their potential function is revealed by comparing wings either with or without the owl's specializations in fluid-mechanical experiments.

Key Research Findings

Owl Wings

Owl wings differ from those of other birds. Figure 1a shows an example wing of a barn owl (*Tyto alba*). The large wing area, in combination with a relatively low body mass, results in a low wing loading allowing the owl to fly slowly and to carry large prey. Specific camber and thickness distributions (Fig. 1b, c) are also characteristic for owl wings [6]. While the proximal wing is highly cambered and anteriorly thickened, the distal wing is thin and less cambered. Anatomical studies revealed the construction plan of the wing (Fig. 1d). Skeletal elements, the muscle mass, and the coverts are responsible for the thick anterior part of the wing. By contrast, the distal wing and the posterior arm part of the wing are formed by remiges. These areas are extremely thin and lightweight. Hand and arm part of the wing have different functions (lift production and thrust) expressed by different geometries and profiling. At both wing parts, however, the airflow tends to separate in wind tunnel experiments when tested with a smooth surface without any surface or edge modifications [5, 16]. Owls, however, evolved microstructures on their feathers to influence the

airflow around the wing in such a way that the airflow stays attached especially during critical flight maneuvers [5, 16] such as takeoff, landing, or striking. Hence, the wings produce a huge amount of lift even at low flight speeds.

Owl Feathers

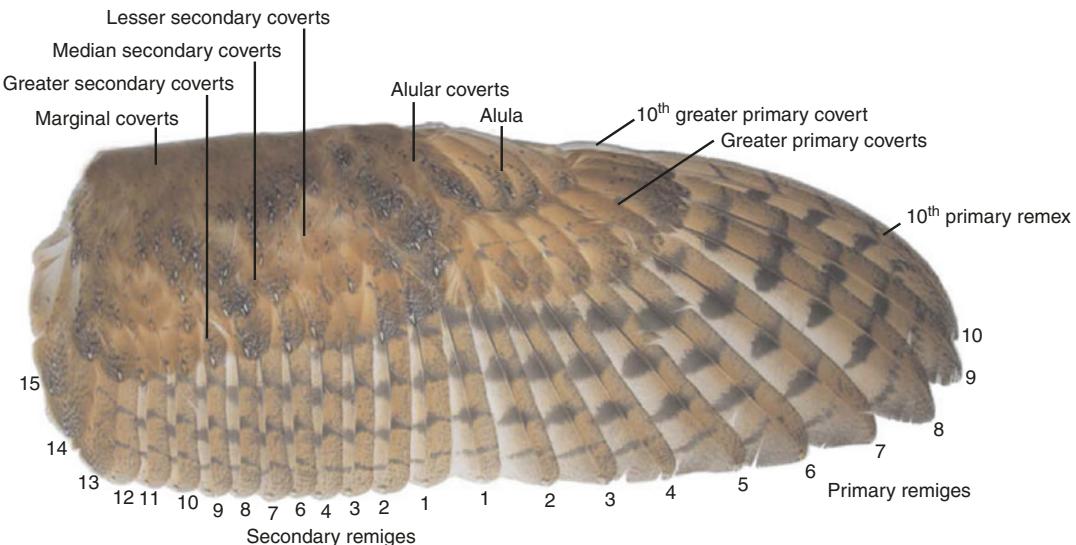
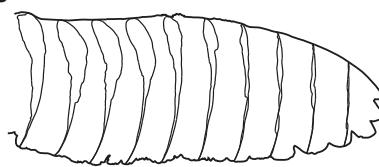
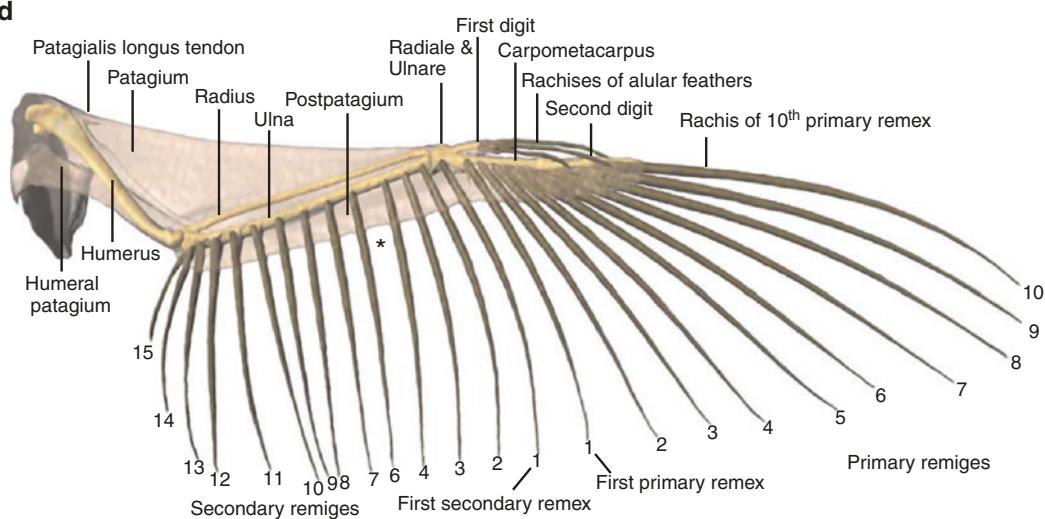
Owl feathers share the anatomy of typical contour feathers [8]. However, they differ in some detail. As feathers are the main aerodynamic components of the bird wing, their geometry and dimensions affect the overall wing geometry. In owls, the feather vanes are large and at the same time extremely light and flexible. Fewer interlinks of the barbs cause these effects [8]. Hereby, owl feathers are very air-transmissive and flexible, whereby they can react rapidly to varying airflow conditions. Their bending behavior is passively driven. It depends on the material properties of feather keratin and the geometry of the rachis and the barbs, respectively [17].

Compared to other birds, owls stand out due to a large count of contour feathers [9]. Specifically, head, feet, and body are densely covered. This upholstery of the body serves the suppression of noise to some extent, but also thermal insulation since owls have only small fat reserves. The coloration of the plumage is mostly adapted to the environment of the relevant owl species with respect to an effective camouflage, but also to sexual display to some extent. Snowy owls (*Bubo scandiacus*), for instance, have a white plumage, while species that live in wooded environment have a brownish patterned plumage (see Fig. 2). Different color intensities between males and females are found, for instance, in barn owls with the females being slightly darker.

The most conspicuous attribute of owl remiges are several anatomical microstructures that are responsible for the silent flight [4, 8]. Interestingly, fishing owls like the Malay fish owl (*Bubo ketupu*) lack such feather specialization [4]. Fish cannot hear the approaching owl. Hence, these owls do not need to fly silently.

Microstructures of Owl Feathers

Owls evolved several surface and edge modifications of their wings that are in turn specializations

a**b****c****d****S**

Silent Owl Wings, Fig. 1 Anatomy of a barn owl (*Tyto alba*) wing: (a) Topography of the dorsal wing (photograph). (b) Surface image of the digitized wing (surface scan). (c) Profiles of the owl wing in steps of

10 % of the half wing span. (d) 3D reconstruction of skeletal elements, skin, and rachises of the remiges (processed CT scan). *The fifth secondary remex is missing in all owl species

of the plumage. Three structures are mainly worth mentioning: the leading-edge comb-like serrations, the velvet-like upper surface, and the inner vane fringes [4, 7, 8].

Each feather that functions as a leading edge of the wing is equipped with comb-like serrations (Fig. 3). The number of remiges forming the leading edge varies among owl species, but is in the

Silent Owl Wings,

Fig. 2 Feathers of different owl species: (a) Tenth primary remex of a barn owl (*Tyto alba*) wing. The outer vane is equipped with tiny serrations, the inner vane is lined with fringes. (b) Contour feather from the body of a snowy owl (*Bubo scandiacus*). Note the symmetry of the vanes in comparison to the flight feathers (a) and (c) and the elongated plumulaceous barbs (fringes at the end of the vanes). (c) Primary remex of an eagle owl (*Bubo bubo*). As the eagle owl is the largest owl, so are its feathers. Scale bars represent 2.5 cm

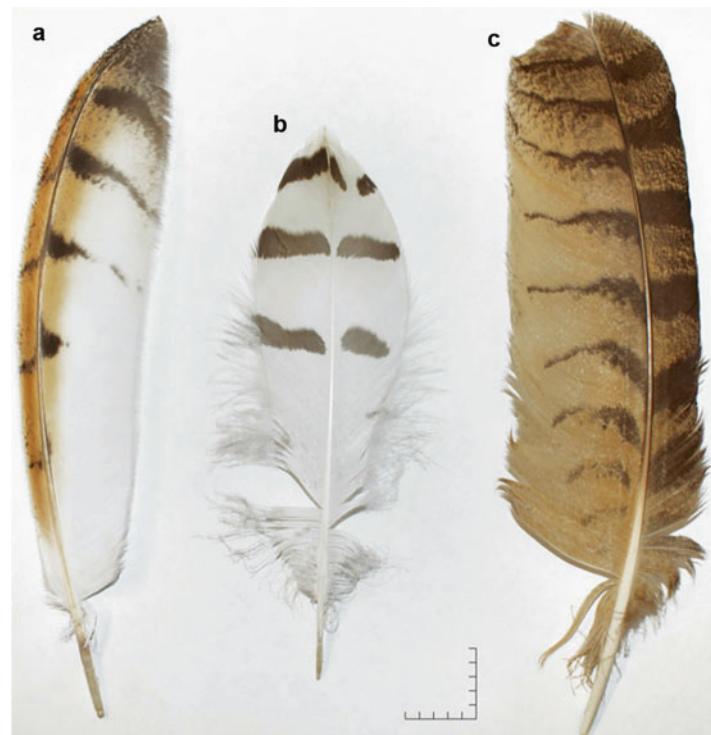
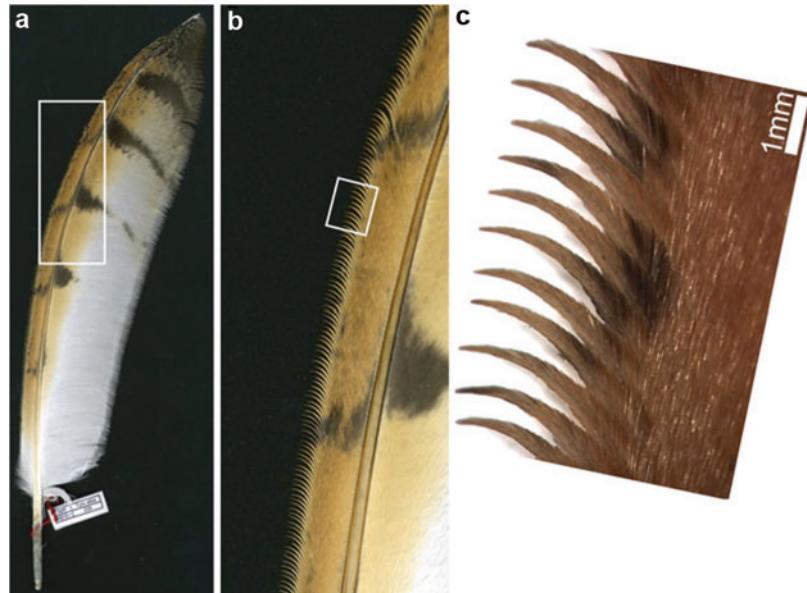
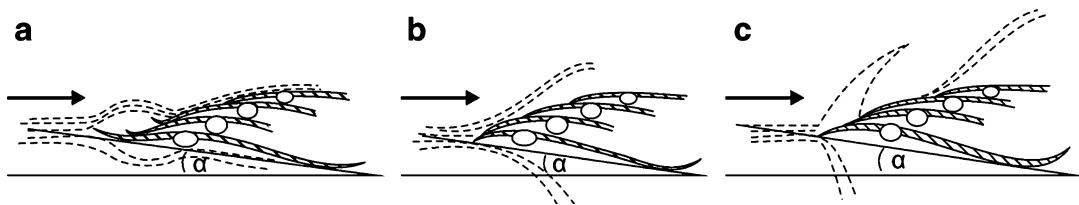
**Silent Owl Wings,**

Fig. 3 Serrations at the leading edge of a tenth primary of a barn owl (*Tyto alba*) in different magnifications: Barb endings separate and bow toward the *upper side* (a, b, c). Note also the fringes at the inner vane (a) and the dorsal surface texture



range of 1–3 remiges. Barb endings separate and bend upward, changing their form and function. Differences in the anatomical phenotype are small within one feather, though small variations occur

(Fig. 3) [8]. While long serrations with small spacing are found in basal and central regions of the feathers, shorter serrations with a larger distance are located at the tip of the feather.



Silent Owl Wings, Fig. 4 Scheme of the airflow at the leading edge of a distal bird wing: (a) Intact wing of a tawny owl (*Strix aluco*) in the region of the serrations. Note the laminar flow conditions around the wing. (b) After manual removing of the serrations at the wing of the

tawny owl, the flow separates early. (c) Flow conditions around a wing of a mallard duck (*Anas platyrhynchos*). The arrow indicates the flow direction; alpha is the angle of attack (Drawings after Neuhaus et al. [5])

Functionally, serrations affect the airflow and noise production only marginally during cruise flight [5]. The stagnation pressure at the leading edge of the wing prevents a functional airflow around the serrations. However, in critical flight maneuvers, such as landing or striking, the angle of attack is much higher and the serrations cause the airflow to stay attached to the wing. This is achieved by dividing the airflow into many small micro-vortices that extent chordwise over the upper wing surface. As a consequence, the boundary layer is more energetic and flow separation is prevented (Fig. 4a). By this means, lift production is maintained even at high angles of attack and low flight speed. This effect gets lost after manual removal of the serrations at the leading edge (Fig. 4b). Here, the flow field around the wing resembles that of a mallard duck (Fig. 4c) which produces much more noise [5].

The second specialization is found on the upper wing surface. Each feather of the owl's plumage is covered with a velvet-like dorsal texture (Fig. 5). This structure is formed by elongations of the hook radiate. Each hook radiate terminates in a filament structure, termed pennula. Due to their large number, the surface becomes very fluffy and porous. Functionally, two aspects are at least associated with this structure.

On the one hand, the velvet-like dorsal surface of the inner vane, which is mostly covered by the adjacent feather, is very thick and well developed. The pennula are long and interconnect at large angles with the feather's surface. This in turn causes a friction-reductive device on the upper surface and enables a smooth and silent gliding

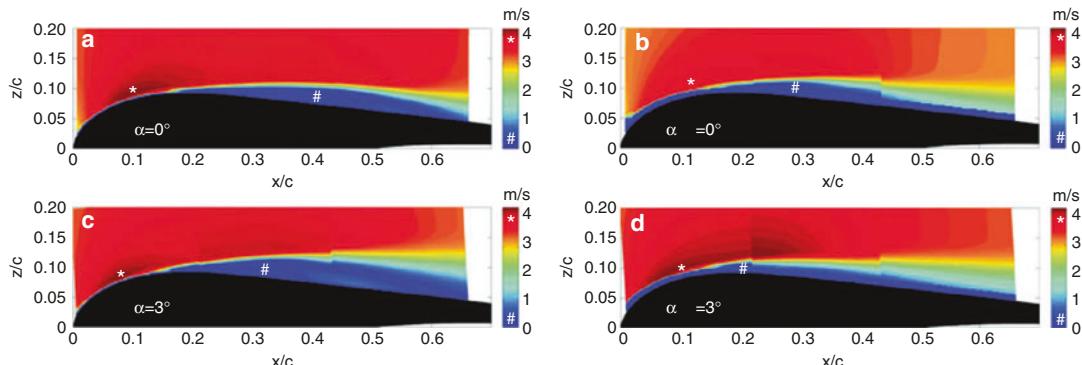
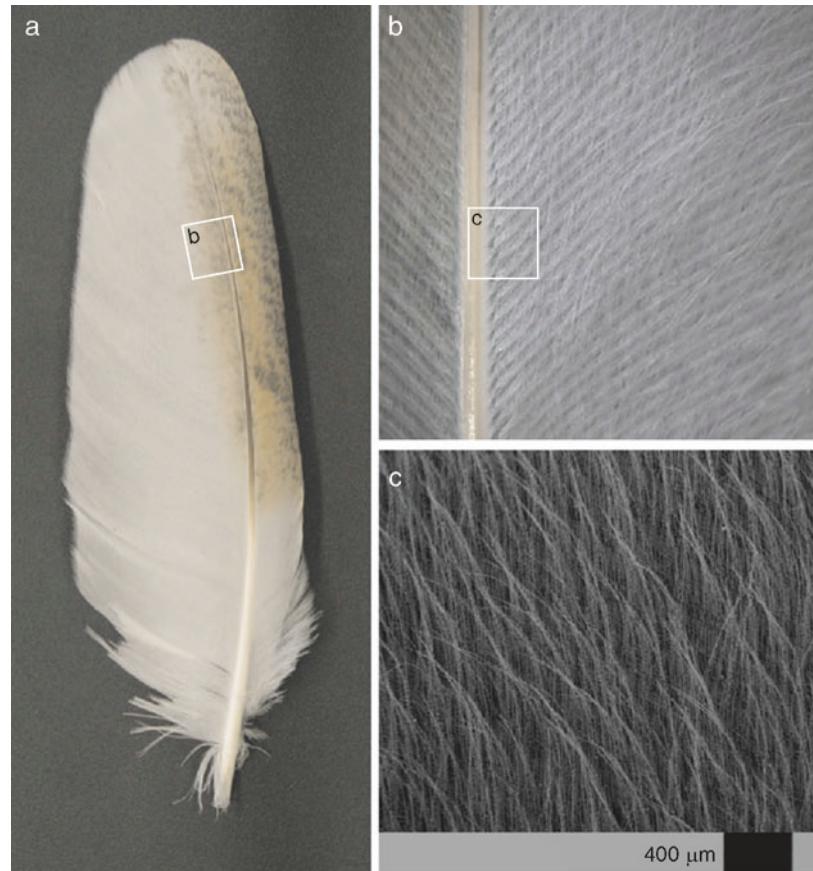
of the feathers. In all other birds investigated so far, the parallel-oriented barbs of adjacent feathers rub against each other producing a well-noticeable high-frequency sound, especially during flapping flight. In owls, however, such noises are prevented by the velvety surface. Those noises that still occur are damped by the porous feather texture acting like acoustic absorbers.

On the other hand, the velvet-like structure is also found at the outer vane. Here, the texture affects the aerodynamics of the wing. As mentioned above, the airflow tends to separate at wings with the specific camber and thickness distribution of owls and a smooth wing surface [16]. Figure 6 shows the PIV results of the velocity distribution of the flow around an owl-based airfoil at Reynolds number 40,000, two different angles of attack (0° and 3°), and two different surfaces. The velocity of the flow is color coded. At the wing with a clean and smooth surface (Fig. 6a, c), a separation bubble occurs on the suction side that increases with the higher angle of attack and leads to a complete flow separation beyond 3° angle of attack (not shown) [16].

Increasing the surface roughness as it is found in owls is one means in the direction of flow control. An artificial velvety surface texture applied to the owl-based wing geometry has a dramatic influence on the flow field (Fig. 6b, d). The velvety surface shifts the separation toward higher angles of attack indicated by a smaller separation bubble in comparison to the clean wing model. The boundary layer of the wing is controlled in such a way that occurring separation bubbles are reduced in size and shifted toward the

Silent Owl Wings,

Fig. 5 Velvet-like dorsal surface of a barn owl's fourth secondary in different magnification: Elongations of the hook radiates create this filigree texture which is supposed to influence the airflow and to reduce friction noise. (a, b) Photographs, (c) scanning electron microscope image

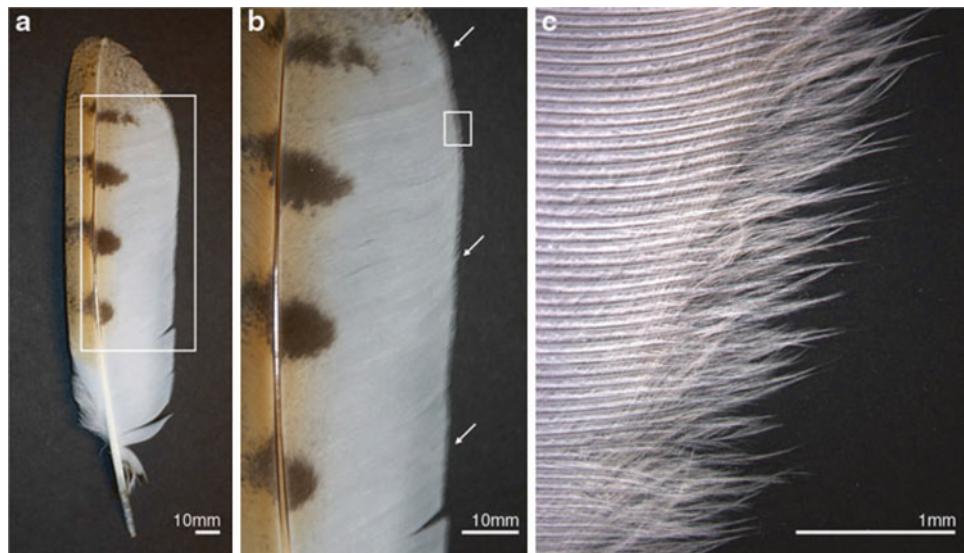


Silent Owl Wings, Fig. 6 Averaged velocity fields at $Re = 40,000$ and an angle of attack of 0° and 3° for a clean wing (a, c) and the same wing geometry with a velvety surface (b, d). The velocity of the airflow is color coded (online version)

leading edge of the wing. Consequently, the airflow is stabilized even at low flight speeds [16].

A further specialization can be found at the inner vanes of the remiges. Here, barb endings

separate by a reduction of the hooklets of the hook radiates. Barbs are no longer interconnected and fringes are formed (Fig. 7). Since the tips of the inner vanes of remiges form



Silent Owl Wings, Fig. 7 Fringes at the inner vane of a barn owl's fifth primary in different magnification: Hooklets of the hook radiates are absent causing the barbs to

separate. Hook and bow radiates change their length and thereby support the formation of fringes. (a) and (b) dorsal, (c) ventral

the trailing edge of the wing, fringes are also found here. The appearance of fringes is very fluffy and unstable in shape. They are made up of three components: the central barb shaft and the laterally attached hook and bow radiates. The radiates support the formation of fringes by a parallel orientation to the barb shafts [8]. Since the barb shaft does not change its shape, as is the case for the serrations, the structure is pliant and the fringes can float freely resulting in a thin, almost two-dimensional structure. Functionally, fringes are assumed to decrease the sound intensity by a reduction of turbulence at the trailing edge of the wing [7, 18]. While sharp edges of wings and flat plates are known to produce a well-noticeable noise even at low Reynolds numbers [19], fringed trailing edges of owl remiges prevent such phenomena. Each wing feather has its own trailing edge that produces noise during flight independently from being part of the wing or separated, for example, during flapping flight. In general, fringes act as a pressure release mechanism that interrupt scattering phenomena which can be found when turbulent eddies pass over the posterior wing region that gradually transitions from the loaded wing to the freestream

conditions [18, 19]. Without such fringes, turbulent boundary layers being convected past the trailing edge into the wake would generate an intense, broadband scattering noise. In owls, fringes prevent sharp edges at the trailing edge such that the turbulent boundary layer is merged smoothly into the airflow.

Behavioral Adaptations

Nocturnal owls are unable to benefit from thermal upwind like other birds of prey. Hence, owls combine active flapping flight and gliding phases [9, 10]. The large wing area and the specific profiling of the wing enable the owl to glide and maneuver with little movements of the wing. The flight speed is relatively low and depends on the illumination of the environment. Barn owls (*Tyto alba*), for instance, can reduce their flight speed down to 2.5 m/s [9, 10] which per se leads to a silent flight, since only little turbulences occur at that velocity. Apart from the slow gliding flight, some adaptations in the motion of the owl already prevent rising noises. During hunting, owls reduce their wing beat frequency and amplitude [9]. Consequently, less friction between feathers occurs during flight which would cause

rubbing noises. Those noises that still rise are efficiently damped by the porous upper wing surface [4, 7].

Future Directions for Research

Urbanization and the increasing demand for air freight and renewable energy have led to a dramatic increase of noise pollution in the last decades. At ground level, airplanes produce an intense noise by their engines and high-lift configurations of the wings. Furthermore, wind farms are constantly upgraded. Their noise emission does not only annoy humans, but also disturbs migratory animals such as birds, or in case of offshore wind farms, whales, and fishes. How flight noise can efficiently be reduced is demonstrated by owls. These birds evolved noise-reduction and noise-suppression devices in the course of evolution. The implementation of the underlying mechanisms in modern aircraft and rotors of wind-power plants would help to reduce noise pollution. Until then more studies and experiments have to be carried out. In joint biological and fluid-mechanical efforts, the basic aerodynamic characteristics and the fluid-mechanical details that define the drag and noise reduction mechanisms of owl wings have to be investigated in more detail. This will be done using preserved material from zoological collections, experiments with living owls, and wing models in wind tunnel experiments. The knowledge gained will be used to create devices that operate at higher Reynolds numbers, since copying the natural structure alone would not deliver satisfying results. In the long run, designing silent wings of modern aircraft will bring us closer to the goal of a quieter environment.

Cross-References

- [Biomimetic Flow Sensors](#)
- [Confocal Laser Scanning Microscopy](#)
- [Friction-Reducing Sandfish Skin](#)

- [Insect Flight and Micro Air Vehicles \(MAVs\)](#)
- [Scanning Electron Microscopy](#)
- [Shark Skin Drag Reduction](#)

References

1. Pennycuik, C.J.: Wing beat frequency of birds in steady cruising flight: new data and improved predictions. *J. Exp. Biol.* **199**, 1613–1618 (1996)
2. Rayner, J.M.V.: On aerodynamics and the energetics of vertebrate flapping flight. *Contemp. Math.* **141**, 351–400 (1993)
3. Rüppell, G.: *Vogelflug*. Rowohlt Taschenbuchverlag GmbH, Hamburg (1980)
4. Graham, T.: The silent flight of owls. *J. Roy. Aero. Soc.* **38**, 837–843 (1934)
5. Neuhaus, W., Bretting, H., Schweizer, B.: Morphologische und funktionelle Untersuchungen über den lautlosen Flug der Eule (*Strix aluco*) im Vergleich zum Flug der Ente (*Anas platyrhynchos*). *Biol. Zentr. Bl.* **92**, 495–512 (1973)
6. Nachtigall, W., Klimbingat, A.: Messungen der Flügelgeometrie mit der Profilkamm-Methode und geometrische Flügelkennzeichnung einheimischer Eulen. In: Nachtigall, W. (ed.) *Biona Report 3*, pp. 45–86. Gustav Fischer Verlag, Stuttgart/New York (1985)
7. Lilley, G.: A study of the silent flight of the owl. *AIAA Paper*, pp. 98–2340 (1998)
8. Bachmann, T., Klän, S., Baumgartner, W., Klaas, M., Schröder, W., Wagner, H.: Morphometric characterisation of wing feathers of the barn owl (*Tyto alba*) and the pigeon (*Columba livia*). *Front. Zool.* **4**, 23 (2007)
9. Mebs, T., Scherzinger, W.: *Die Eulen Europas*. Franckh-Kosmos, Stuttgart (2000)
10. Taylor, I.: *Barn Owls: Predator – Prey Relationship and Conservation*. Cambridge University Press, Cambridge (1994)
11. Knudsen, E.: The hearing of the barn owl. *Sci. Am.* **245**(69), 113–125 (1981)
12. Hausmann, L., Campenhausen, V.M., Endler, F., Singheiser, M., Wagner, H.: Improvements of sound localization abilities by the facial ruff of the barn owl (*Tyto alba*) as demonstrated by virtual ruff removal. *PLoS One* **4**(11), e7721 (2009)
13. Konishi, M., Takahashi, T., Wagner, H., Sullivan, W., Carr, C.: Neurophysiological and anatomical substrates of sound localization in the owl. In: *Auditory Function – Neurobiological Bases of Hearing*, pp. 721–745. Wiley, New York (1988)
14. Konishi, M.: How the owl tracks its prey. *Sci. Am.* **61**, 414–424 (1973)
15. Sarradj, E., Fritzsch, C., Geyer, T.: Silent owl flight: bird flyover noise measurements. *AIAA Paper 2010-3991* (2010)

16. Klän, S., Bachmann, T., Klaas, M., Wagner, H., Schröder, W.: Experimental analysis of the flow field over a novel owl based airfoil. *Exp. Fluids* **46**, 975–989 (2008)
17. Bonser, R.H., Purslow, P.: The young's modulus of feather keratin. *J. Exp. Biol.* **198**, 1029–1033 (1995)
18. Herr, M.: Experimental study on noise reduction through trailing edge brushes. *Note. N. Fl. Mech. Mul. D* **92**, 365–372 (2006)
19. Moreau, D.J., Brooks, L.A., Doolan, C.J.: On the aeroacoustic tonal noise generation mechanism of a sharp-edged plate. *JASA E. L* (2011). doi:10.1121/1.3565472

Simulating Nanoscale Heat Transport

Giuseppe Romano¹, Jean-Philippe M. Peraud²
and Jeffrey C. Grossman¹

¹Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

²Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Silica Gel Processing

- Sol–Gel Method

Silicon Microphone

- Electrostatic MEMS Microphones

Silicon Thin Films

- Heat Transfer in Semiconductor Nanostructures

Silks of Spiders as Model Bio-polymers

- Spider Silk

Silver (Ag)

- In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles

Synonyms

- Modeling nanoscale heat transport

Introduction

Heat conduction has been modeled for almost two centuries by the well-known Fourier's law. Jean-Baptiste Joseph Fourier stated that “the quantity of heat which flows uniformly, during unit of time, across unit of surface taken on any section whatever parallel to the sides, all other things being equal, is directly proportional to the difference of the extreme temperatures, and inversely proportional to the distance which separates these sides” [1]. Fourier's law can be conveniently written in its local form:

$$\mathbf{J} = -\kappa \nabla T, \quad (1)$$

where κ is the thermal conductivity (for silicon, it is about $150 \text{ W m}^{-1} \text{ K}^{-1}$), T is the lattice temperature, and \mathbf{J} is the heat flux. Equation 1 together with the continuity equation for thermal flux

$$\rho C_p \frac{\partial T}{\partial t} + \nabla \cdot \mathbf{J} = H, \quad (2)$$

gives the well-known diffusive heat conduction equation. In Eq. 2, C_p is the specific heat capacity, ρ the material density, and H the sum of all the heat sources in the system.

While Fourier's law is valid at the macroscale, it has been experimentally proved that in some nanostructured materials, such as nanowires [2], thin films [3], or nanoporous materials [4], the description of thermal transport as diffusive heat conduction breaks down. Furthermore, as we will see later, Fourier's law is violated in transient heat conduction experiments [5, 6].

The departure of heat transport from the diffusive regime can be qualitatively explained by considering heat as being carried by material waves, with their own frequency, group velocity, and dispersion curve [7]. Wave effects can be taken into account by the so-called molecular dynamics (MD) simulations, which essentially apply Newton's law to a system of atoms, whose internal forces are modeled by either empirical or first-principles calculations. The thermal conductivity computed by a MD-based formalism inherently includes both harmonic and anharmonic effects and can be computed by the well-known Green-Kubo formula:

$$\kappa_{\alpha\beta} = \frac{1}{V k_B T^2} \int_0^\infty \langle J_\alpha(0) J_\beta(t) \rangle dt \quad (3)$$

where k_B is the Boltzmann constant, V the volume of the system, and $\kappa_{\alpha\beta}$ the thermal conductivity tensor. In Eq. 3, $\langle A \rangle$ is the average ensemble of the observable A and can be replaced by a time average if the simulation is long enough to satisfy ergodicity [8]. Although MD is a powerful approach to model thermal conductivity at very small scales, it can become computationally prohibitive when dealing with realistic structures of mesoscale dimensions, such as thin films and porous materials. To overcome this limit, one may model the quantized lattice vibrations – or phonons – as particles and compute the temperature as well as the thermal fluxes by applying numerical schemes solving for particle transport problems.

This simplification is justified by noting that at room temperature, the dominant phonons in certain materials, such as Si, have wavelengths below 10 nm. Within this assumption, phonon transport can be safely modeled by the Boltzmann transport

equation (BTE) [9]. The BTE has been gaining much attention recently in modeling thermal transport in nanostructured materials, especially for thermoelectric applications. In the following, we will focus on the BTE and the two main approaches used to solve it in complex structures. Readers interested in modeling wave effects at the mesoscale can refer to [10]. Also, while here we discuss the potential of the BTE in nanoscale heat transport simulations, there are many equally important methods left aside. A comprehensive review on computational tools for heat transfer can be found in [11].

The Boltzmann Transport Equation

In order to introduce the BTE for phonons, we first define $f(\mathbf{r}, \mathbf{k}, t)$ as the phonon distribution function at a given time t , at position \mathbf{r} with wave vector \mathbf{k} . With no loss of generality, we assume the Brillouin zone to be isotropic. We further define the quantity $I(\mathbf{r}, \mathbf{s}, \omega, p) = \frac{1}{4\pi} |\mathbf{v}(\omega, p)| f D(\omega, p) \hbar \omega$, where $D(\omega, p)$ is the density of states, $\mathbf{v}(\omega, p)$ the group velocity, $\hbar \omega$ the phonon energy, and p the phonon branch. The group velocity can be obtained from the dispersion relations $\omega(\mathbf{k})$ as $\mathbf{v} = \frac{\partial \omega(\mathbf{k})}{\partial \mathbf{k}}$. The quantity $I(\mathbf{r}, \mathbf{s}, \omega, p)$ represents the intensity of phonons for a given frequency within a unit solid angle [12], traveling along the direction $\mathbf{s} = \frac{\mathbf{v}}{|\mathbf{v}|}$. In the absence of any external driving forces, such as an applied temperature gradient, the intensity $I(\mathbf{r}, \mathbf{s}, \omega, p)$ equals the equilibrium intensity $I_0(T_0)$, defined as $I_0(T_0) = \frac{1}{4\pi} |\mathbf{v}(\omega, p)| f_0(\omega, T_0) D(\omega, p) \hbar \omega$. The term $f_0(T_0)$ is the Bose-Einstein distribution $f_0(T_0) = \left[\exp\left(\frac{\hbar \omega}{k_B T_0}\right) - 1 \right]^{-1}$. Assuming that all phonon scattering events are uncorrelated, the nonequilibrium phonon intensity can be modeled by

$$\frac{1}{|\mathbf{v}|} \frac{\partial I}{\partial t} + \mathbf{s} \cdot \nabla I = \frac{I_0(T) - I}{\tau |\mathbf{v}|}, \quad (4)$$

which is the BTE under the so-called relaxation time approximation [12]. In Eq. 4, τ is the intrinsic scattering time [13]. Energy conservation among

modes can be obtained by applying the continuity equation for the energy flux, which results in

$$\sum_p \int_0^{\omega_M^p} \frac{I_0(T)}{\tau} d\omega = \sum_p \int_0^{\omega_M^p} \frac{ < I > }{\tau} d\omega, \quad (5)$$

where $I_0(T)$ is the Bose-Einstein distribution at the temperature T , ω is the phonon angular frequency, p is the phonon branch, and ω_M^p is the maximum angular frequency for the branch p . In practical thermal conductivity calculations, a difference of temperature ΔT is applied to a simulation domain, and once Eqs. 4 and 5 are solved consistently, the thermal flux is computed by

$$\mathbf{J}(\mathbf{r}) = 4\pi \sum_p \int_0^{\omega_M^p} |\mathbf{v}| < \mathbf{Is} > d\omega, \quad (6)$$

and the *effective* thermal conductivity is deduced by $J = -\kappa_{eff} \nabla T$. The BTE described in Eq. 4 has to be solved for the whole phonon spectrum and is called the *frequency-dependent* BTE. Generally speaking, the BTE can be solved either deterministically or stochastically. In the following, we devote a section to each approach.

Deterministic Solution of the BTE

Here we show how the BTE can be useful for calculating the steady-state thermal conductivity values in nanostructured materials. Following the approach described in [14], we consider only steady-state transport. Furthermore, we assume that the bulk thermal conductivity is isotropic and that a very small temperature difference is applied across the sample. Under these simplifications, the BTE becomes

$$\mathbf{As} \cdot \nabla \tilde{T} + \tilde{T} = \gamma \int_0^{\infty} \frac{K}{\Lambda'^2} < \tilde{T} > d\Lambda', \quad (7)$$

where \tilde{T} is the departure of a temperature associated with a given phonon mode from the equilibrium, normalized by the applied temperature

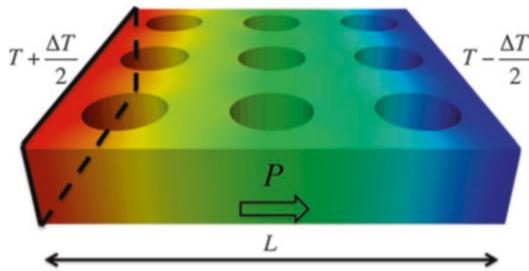
difference. In Eq. 7, the term $K(\Lambda)$ is the bulk phonon mean free path distribution, a quantity that can be obtained either theoretically [8] or experimentally [15, 16], and $\gamma = \left[\sum_p \int_0^{\infty} \frac{K}{\Lambda^2} d\Lambda \right]^{-1}$ is a material. The notation $< x >$ stands for the angular average $< x > = \frac{1}{4\pi} \int_{4\pi} x d\Omega$. The right-hand side of Eq. 7 is related to the normalized *effective* lattice temperature.

Practically, Eq. 7 requires the discretization of $K(\Lambda)$, as opposed to the discretization of the phonon frequencies typically used in a frequency-dependent approach. For each MFP, the BTE is solved by means of the finite-volume method, whereas the solid angle is discretized by the discrete ordinate method [17]. Equation 7 is named phonon MFP-BTE and has the following advantages:

- It retains the accuracy of the frequency-dependent approach.
- It requires the knowledge of only the bulk MFP distribution, which is a quantity that can be obtained from experiments [15].
- The requirements in the discretization of the bulk MFP distribution are less demanding than in a typical frequency-dependent approach, leading to a significant improvement in the computational efficiency [14].
- It is relatively easy to parallelize. In fact, each phonon mode is only coupled through the integral appearing on the right-hand side of Eq. 7. In a typical iterative solver, each BTE for a given MFP can be run independently, using the effective lattice temperature from the previous step.

In the following, we show an application of the MFP-BTE to nanoporous silicon, a promising material for thermoelectric applications, thanks to its capability to suppress thermal transport with little degradation in the electrical conductivity [4].

As shown in Fig. 1, a difference of temperature ΔT is applied to the simulation domain. Then, after the MFP-BTE converges, the thermal power, P , is computed on either the cold or hot



Simulating Nanoscale Heat Transport,

Fig. 1 Simulation of a porous material. A difference of temperature ΔT is applied across the two ends of the domain

side. The effective thermal conductivity κ_{eff} is computed by using Fourier's law $\kappa_{\text{eff}} = \frac{PL}{A\Delta T}$, where A is the contact area. In Fig. 2a, the discretization of the solid angle is shown. As the actual simulated system is two-dimensional, we only consider the upper hemisphere and then apply symmetry. Figure 2b shows the discretization of the spatial domain. The MFP distribution, shown in Fig. 2c, is obtained by means of a frequency-dependent model, as described in [14].

The phonon thermal conductivity (PTC) is computed by varying the length of the unit cell from the nanoscale to the macroscale. As shown in Fig. 3, for very small unit cells, the thermal conductivity is reduced by roughly seven times with respect to the diffusive value, which is given by the approximated formula [18]

$$\kappa_{\text{eff}} = \kappa_{\text{bulk}} \frac{1 - \phi}{1 + \phi}, \quad (8)$$

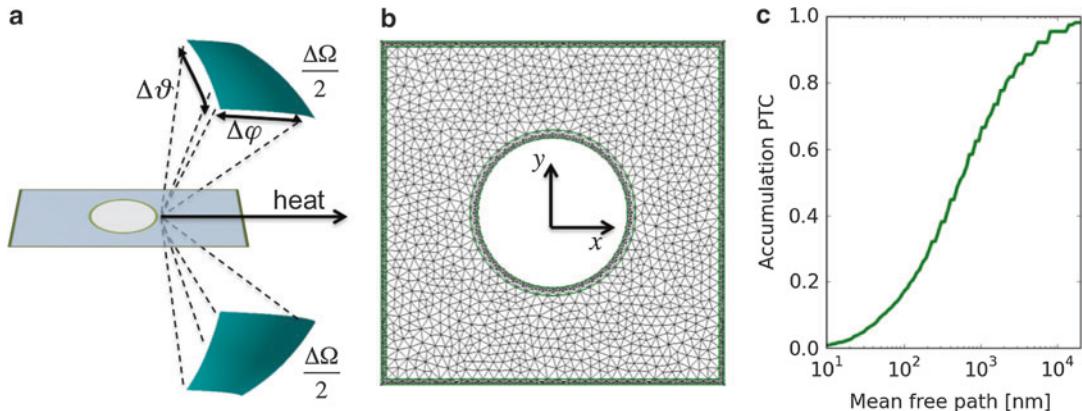
where ϕ is the material porosity. In our case, we choose $\phi = 0.25$, which leads to the PTC $\kappa_{\text{eff}} \approx 77 \text{ W/mK}$. For very large unit cells, the PTC correctly recovers the diffusive limit. In our calculations, the pore walls diffusively scatter phonons. In general, depending on the surface roughness, phonon wavelength, and temperature, there might be a fraction of phonons specularly reflected [9]. The BTE-MFP has also been applied to the realistic case described in [4], finding good agreement with experiments. These calculations

represent a validation of the developed code and a starting point for PTC minimization in porous materials with different pore configurations. For example, in [19], it has been shown that triangular pores arranged in misaligned columns bring a reduction in the PTC of about 60 % with respect to the aligned case with the same porosity.

Monte Carlo Methods

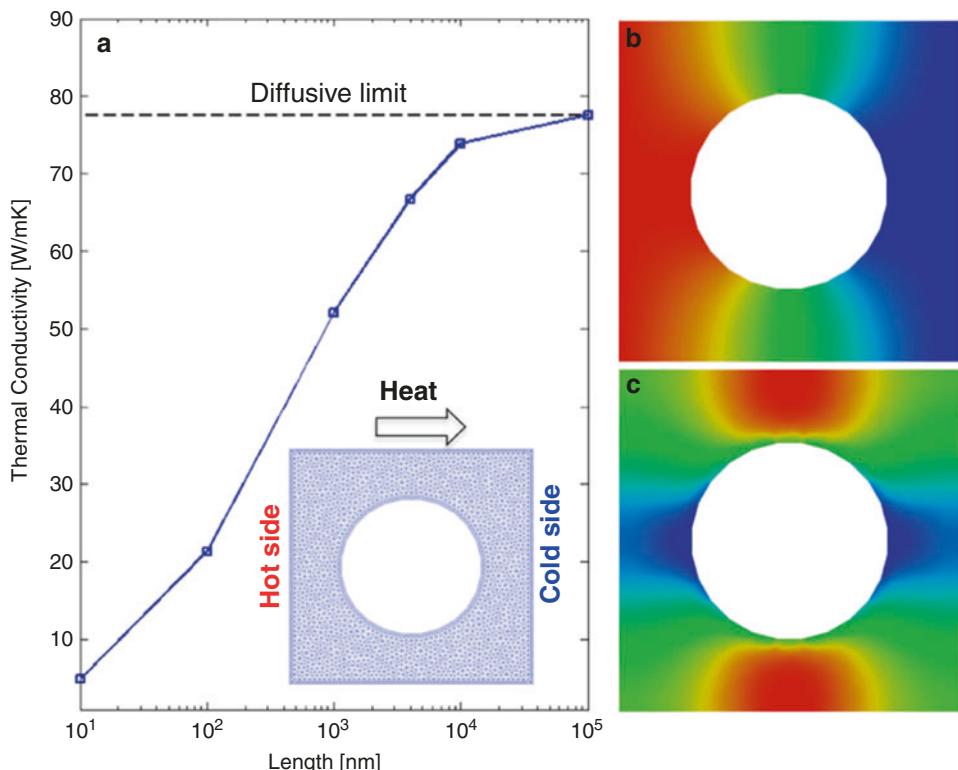
One of the major limitations of traditional deterministic solution methods lies in the discretization of the six dimensions – seven dimensions for transient problems – of the phase space. Particle Monte Carlo methods have gained popularity thanks, in part, to their ability to avoid the full discretization of the phase space. They are commonly used for solving the Boltzmann equation for a range of physics problems such as neutron transport, radiation, rarefied gases, or electron transport. The idea of using such an approach for solving phonon transport problems was first introduced by Klitsner [20] and further improved by Peterson [21] and Mazumder et al. [22]. Various improvements were added between the years 2000 and 2014 and made it a powerful method.

Monte Carlo methods for particle transport solve the Boltzmann equation by *directly simulating* the physical processes involved. Phonons are characterized by their position, their frequency, their polarization, and their traveling direction, although alternative descriptions, using, for instance, the wave vectors instead of the frequencies, are also adequate. The group velocity vector is deduced from these properties, using the dispersion relation. Simulated phonon properties are randomly drawn from specified distributions in the phase space, and the rules for their evolution in time are deduced from the physical formulation. Quantities of interest, such as temperature or heat flux, are calculated by ensemble averaging. The Monte Carlo algorithm that solves for the transient nonlinear BTE with the relaxation time approximation starts with an initialization step where the initial properties of a population of N computational phonons are drawn from a known distribution defined by the initial



Simulating Nanoscale Heat Transport, Fig. 2 (a) Uniform scheme for solid-angle discretization. (b) Discretization of the spatial domain. (c) Cumulative thermal conductivity for Si obtained by a frequency-dependent

model. We acknowledge ASME for granting us the permission for using Fig. 1 of the article G. Romano and J. C. Grossman. *Journal of Heat Transfer* 137.7 (2015): 071302



Simulating Nanoscale Heat Transport, Fig. 3 (a) Thermal conductivity versus the length of the unit cell. For very large unit cells, thermal transport reaches the diffusive limit. In the inset, the discretization of the simulation domain is shown. Periodic boundary conditions are

applied along the direction of the heat flux. (b) Normalized temperature distribution. (c) Normalized thermal flux distribution. Due to phonon-boundary scattering, heat travels in areas that are far away from pore boundaries

condition. The evolution in time of this population is then calculated through a split algorithm where advection and scattering of phonons are treated separately. The algorithm is described in detail in [22]; further details are provided in [23, 24]. A timestep Δt is defined, and starting from the positions at time t , positions at time $t + \Delta t$ are deduced assuming that particle trajectories are collision-free. This is the advection step. The scattering step consists of simulating the scattering events that occur between times t and $t + \Delta t$. Given the relaxation time τ of a computational particle, the probability of occurrence of a scattering event may be calculated by $1 - \exp(-\Delta t/\tau)$. Thus, this probability law is used to randomly select particles undergoing a scattering event. Selected particles are replaced by new particles whose properties are drawn from the local “post-scattering” distribution $I^0(T)/(\nu\tau)$.

The scattering process must conserve energy. In early work [22–24], this was traditionally done approximately by adding and deleting particles until the post-scattering energy matched the pre-scattering one. Recently, a method for enabling exact energy conservation was proposed [25]. It relies on the simulation of computational particles representing a fixed amount of energy instead of a fixed number of phonons. As a result, energy is rigorously conserved by simply conserving the number of computational particles. Finally, the scattering step requires the knowledge of the local temperature at every timestep. This is usually done by defining a spatial grid of computational cells in order to sample the energy density of the particle population and relating it to the temperature of the corresponding equilibrium (Bose-Einstein) distribution. This algorithm is very similar to the direct simulation Monte Carlo (DSMC) method used for rarefied gases [26].

The main source of error from Monte Carlo methods is the statistical uncertainty, or noise, inversely proportional to the square root of the number of independent samples (the particles) used. Noise is an important issue for problems featuring low deviations from the Bose-Einstein equilibrium since it tends to obscure the signal. Such problems are ubiquitous in nanotechnology and yet cannot be treated by the

particle Monte Carlo method presented above without resorting to massively parallel computing. Recently, a variance reduction scheme, also called “deviational” algorithm, was proposed for rarefied gases [27, 28] and for phonon transport [25] and was shown to capture arbitrarily low deviations from equilibrium at constant computational cost. The method relies on the concept of control variates. Instead of simulating random particles representing the absolute phonon distributions, the variance-reduced algorithm simulates particles representing the *deviation* from a known equilibrium. In other words, if the temperature of the system is expected to feature low deviations from a given temperature T_{eq} , we can introduce the following algebraic decomposition:

$$I = I^0(T_{\text{eq}}) + (I - I^0(T_{\text{eq}})). \quad (9)$$

to obtain the deviational BTE

$$\frac{1}{\nu} \frac{\partial I^d}{\partial t} + \mathbf{s} \cdot \nabla_{\mathbf{x}} I^d = \frac{I^0(T_{\text{loc}}) - I^0(T_{\text{eq}}) - I^d}{\nu\tau}. \quad (10)$$

The deviational algorithm, which solves for $I^d = I - I^0(T_{\text{eq}})$, retains the main features of the non-variance-reduced algorithm and adapts them for solving the deviational BTE. Notably, I^d may be negative and simulated particles must carry a sign. The particles representing the initial condition are drawn from the deviational initial distribution and the advection step of the split algorithm is unchanged. Although the selection rule for scattered particles is also unmodified, the post-scattering properties should now be drawn from the local deviational distribution $(I^0(T_{\text{loc}}) - I^0(T_{\text{eq}}))/\tau$. The total energy density is found by adding the stochastic deviational energy density to the deterministically calculated energy density corresponding to the equilibrium distribution.

The variance-reduced algorithm, discussed in detail in [25], offers the three following advantages:

- It does not introduce any approximation with respect to the traditional algorithm.

- The reduction of the statistical uncertainty essentially results from the fact that the amplitude of the simulated deviational distribution is small with respect to $I^0(T_{\text{eq}})$. In particular, arbitrarily low deviations from equilibrium can be simulated with no additional cost.
- The algebraic decomposition is inherently multiscale. It focuses the calculation effort to the regions where deviation from equilibrium is nonzero. By extending this idea to simulations of deviation from a Fourier solution, one can achieve algorithms which use particles *only* when the deviation from the Fourier solution is nonzero, that is, where size effects are important – the definition of a multiscale method. The simulation of a thermoreflectance experiment presented in [25] is a compelling example of this effect.

Additional computational benefits can be obtained by combining the deviational formulation with the linearization of the BTE, which amounts to linearizing the collision operator as follows:

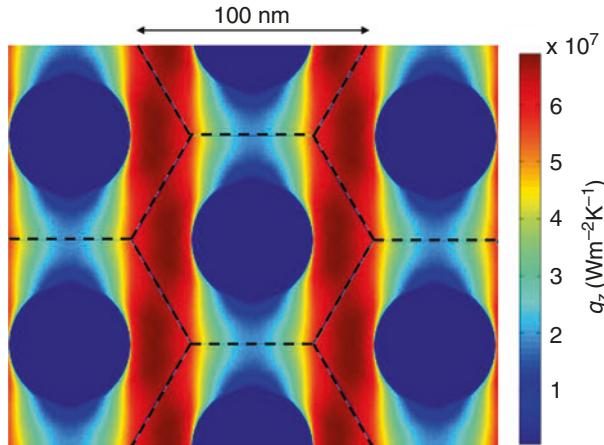
$$\frac{(I^0(T_{\text{loc}}) - I^0(T_{\text{eq}}))}{\tau(\omega, p, T)} \approx \frac{(T_{\text{loc}} - T_{\text{eq}})}{\tau(\omega, p, T_{\text{eq}})} \frac{\partial I^0(T_{\text{eq}})}{\partial T} \quad (11)$$

In other words, the distribution from which the post-scattering properties are drawn does not depend on the local temperature since, after normalization, the temperature term $T_{\text{loc}} - T_{\text{eq}}$ cancels. We recall that the algorithms presented above require the use of a timestep Δt because knowledge of the temperature field (which features in $I^0(T_{\text{loc}})$) is required for simulating the scattering of the particles. Within the linearized approximation, such information is not needed. The scattering rates can be taken at temperature T_{eq} , and the post-scattering properties are all taken from a fixed distribution. Finally, we already explained that energy conservation is ensured by conserving the number of computational particles in the energy-based formulation. These observations yield the following consequences. Particles may be simulated one by

one, independently from one another (a timestep is not needed). For a given particle, the time between each scattering event may be simply calculated from an exponential distribution with survival parameter $\tau(\omega, p, T_{\text{eq}})$. In other words, the traveling time between each scattering event is calculated by the formula $\Delta t = -\tau(\omega, p, T_{\text{eq}}) \ln(R)$, where R is a random number uniformly drawn in the range $(0, 1)$. The resulting “kinetic-type” algorithm is similar to existing algorithms used for neutron or photon transport, where particles are assumed to interact with the underlying medium only. Details on the implementation may be found in Refs. [29, 30]. Reference [31] shows that, at an equilibrium temperature of 300 K, the linearized approximation is reasonably accurate up to a deviation of 30 K.

Although particle methods are inherently explicit in time, it is shown in [29] that most steady problems can be efficiently (and rigorously) treated using this approach. Traditional time-based Monte Carlo algorithms for phonon transport typically solve steady-state problems by letting a time-dependent system evolve toward steady state. Since quantities of interest are then only sampled when the steady state is reached, the number of timesteps needed to compute the transition from the initial condition to the steady state wastes significant computational resources. On the contrary, by treating each particle independently, the “kinetic-type” approach can deliberately ignore the transitory regime and only simulate particles at the steady state. Such particles are emitted from the steady sources only, and particle trajectories are terminated only when they leave the spatial domain, independently of how long the particle has been staying in the system. The resulting method rigorously solves the steady Boltzmann equation. Figure 4 shows an example of steady-state calculation in a nanoporous material.

The kinetic Monte Carlo method yields two immediate advantages. It features substantial memory savings since particles are simulated one by one, and eliminating the need of a timestep significantly reduces the cost of calculating each trajectory. Speedups of a factor 100–1,000 with



Simulating Nanoscale Heat Transport, Fig. 4 A component of the heat flux in a periodic nanostructure with circular pores arranged into a honeycomb structure. The material parameters used for this simulation are the silicon parameters used in [25]. The diameter of the circular pores is 50 nm. The heat flux results from an externally applied temperature gradient of $1 \times 10^6 \text{ Km}^{-1}$ (see Ref. [29] for

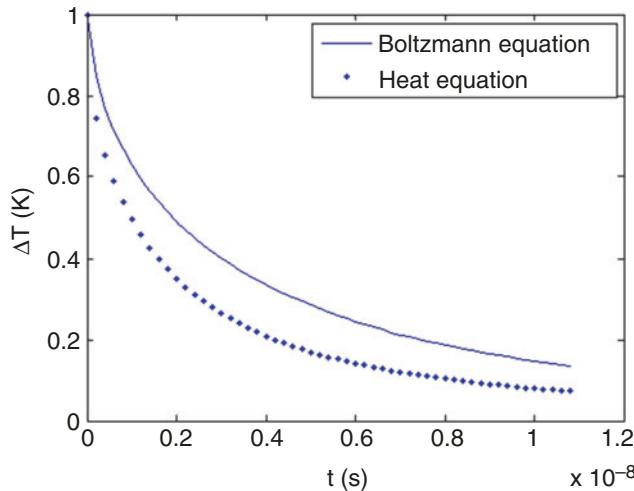
detailed information on the formulation that enables the simulation of an externally applied temperature gradient with periodic boundary conditions). The component of the heat flux shown here is parallel to the applied gradient. In this calculation, the boundaries of the nanopores diffusely reflect the phonons

respect to the timestep-based deviational algorithm have been reported [29]. Finally, the method is inherently multiscale in time. Relaxation times in typical materials such as silicon are known to span several orders of magnitude. Since the time between each scattering process is computed independently of other phonons, the process automatically adapts itself to the time scale of the relaxation time of each step. It should also be noticed that this algorithm completely removes the need for spatial and time discretization. Not only does it remove the associated errors, it also allows to simulate systems of infinite sizes. Figure 5, which shows the time-dependent Boltzmann solution in an infinite system, highlights these features.

An important aspect of particle Monte Carlo methods lies in the fact that they primarily return *moments* of the underlying solution. For instance, the temperature at a given point in space is calculated in an average sense, within a volume surrounding this point. The accuracy of the estimate thus depends on the number of particle trajectories intersecting the volume and contributing to the estimate. Consequently, estimates calculated

within small volumes tend to feature high statistical uncertainties. Recently, the adjoint Boltzmann transport equation for phonons was introduced as a means of addressing this limitation [31]. The adjoint BTE describes a particle problem where particles travel backward in time and where sources and detectors are switched. Thanks to this formulation, estimates can be produced in arbitrarily small volumes in the phase space, including surfaces and points. This is useful, for instance, for producing the contributions of individual phonon modes to the heat flux [32, 33].

The techniques presented above focused exclusively on the relaxation time approximation which, while relatively convenient and reasonably accurate for a number of materials, fails to capture the specific features of three-phonon scattering. The three-phonon scattering operator, which incorporates both energy and momentum conservation, has been widely acknowledged to be a more accurate description [34]. In addition, while the relaxation time approximation is semiempirical and requires the fitting of parameters, the three-phonon scattering may be entirely derived from ab initio calculations.



Simulating Nanoscale Heat Transport, Fig. 5 In this figure, a homogeneous infinite material is considered (the material properties of silicon were used for the calculation). The computational domain is \mathbf{R}^3 . The initial condition is as follows: $T(x, y, z, t = 0) = 301\text{ K}$ within the cube defined by $0 < x < L$, $0 < y < L$, and $0 < z < L$, with $L = 2\text{ }\mu\text{m}$, and $T(x, y, z, t = 0) = 300\text{ K}$ outside of the cubic region. Following this initial heating, the linearized Monte Carlo technique can be used to calculate the average

temperature within the $2\text{ }\mu\text{m}$ cube against time. The deviational distribution is defined with respect to the equilibrium at temperature $T_{\text{eq}} = 300\text{ K}$. The quantity ΔT therefore corresponds to $\Delta T = T - T_{\text{eq}}$. In this configuration, the solution of the heat equation can be calculated analytically. The analytical solution requires the knowledge of the diffusivity, which can be computed from the material parameters

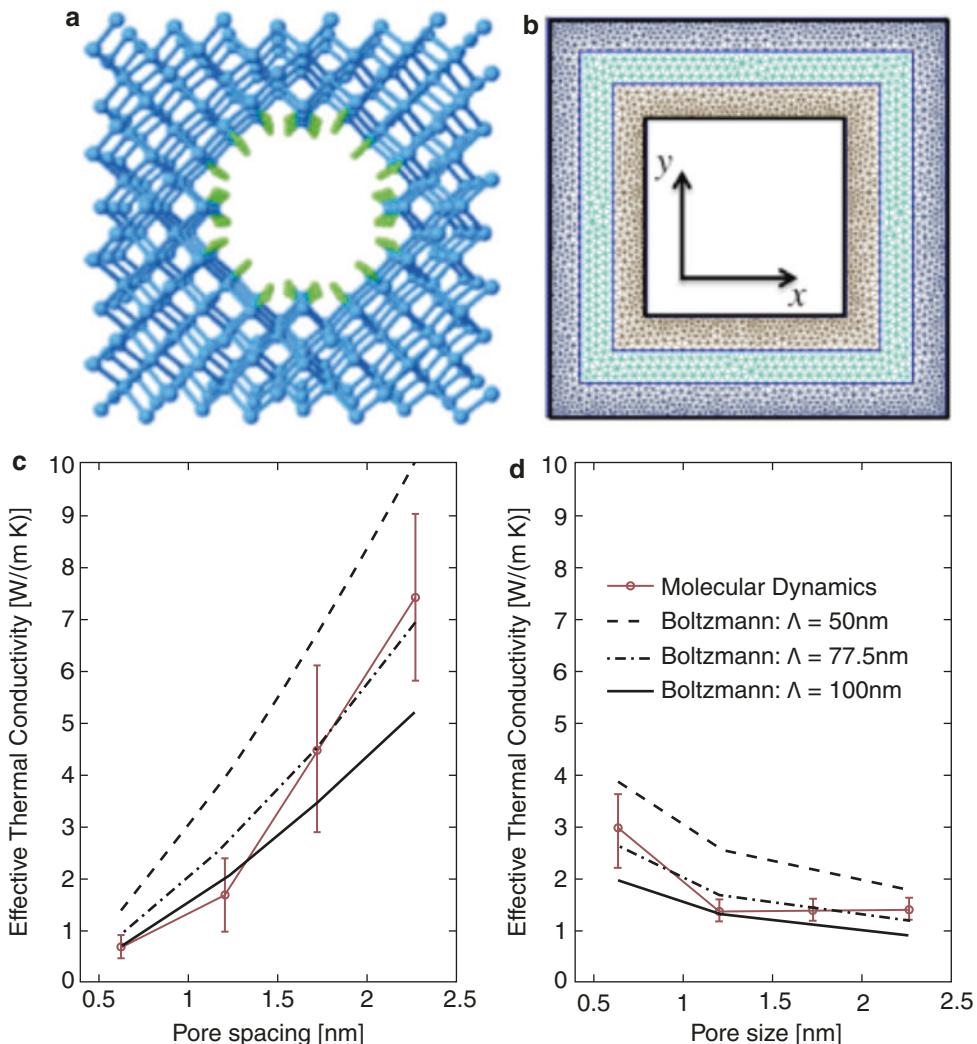
Unfortunately, the resulting BTE is so complicated that very few general methods incorporating three-phonon scattering have so far been developed. Notably, efficient solution techniques for solving the *homogeneous* equation were recently developed [35]. Monte Carlo methods are well known for their ability to solve highly complex partial differential equations and, as such, hold significant potential. Landon and Hadjiconstantinou recently developed a promising solution method based on the deviational approach for treating 2D materials such as graphene [36].

Need for Multiscale Modeling

In the previous sections, we have described two solution techniques for BTE for nano- and mesoscale thermal transport. Also, in the very beginning, we have briefly described the

molecular dynamics method and its importance in certain nanoscale systems such as those dominated by interfaces. In Fig. 6, we report a comparison between the MD and BTE calculations, varying pore distances, and sizes in porous silicon.

Specifically for this case, we used the so-called gray approximation, where all phonons are assumed to have the same MFP, which is used as a parameter to fit MD data. Although MD and BTE results agree qualitatively with each other, the BTE is not as accurate as MD, especially because the phonon dispersion differs from the bulk on very small scales. It is, therefore, paramount for future research to identify the range of validity of both models and ultimately bridge them in accurate yet computationally affordable simulations. In reference to the porous system, when pores are filled with a guest material, interfacial effects may become important and a detailed treatment of phonon transmission



Simulating Nanoscale Heat Transport, Fig. 6 (a) Porous silicon modeled atomistically. The pore walls have been hydrogenated. (b) Porous silicon modeled by

means of a continuum domain. (c) Effective thermal conductivity versus pore spacing. (d) Effective thermal conductivity versus pore size

across such an interface is necessary. On the other hand, performing MD over the whole domain is computationally prohibitive. As a consequence, a good approach could be to split the simulation domain in two parts, one where MD is performed and one governed by the BTE. A good multiscale method should then ensure flux conservation between the two regions and define appropriate rules for the atomistic-to-continuum coupling [37].

Cross-References

- ▶ [Heat Transfer in Semiconductor Nanostructures](#)
- ▶ [Nanostructured Thermoelectric Materials](#)
- ▶ [Thermal Conductivity and Phonon Transport](#)

References

- Fourier, J.B.J.: *Théorie Analytique de la Chaleur*. Cambridge University Press, New York (1822)

2. Li, D., Wu, Y., Kim, P., Shi, L., Yang, P., Majumdar, A.: Thermal conductivity of individual silicon nanowires. *Appl. Phys. Lett.* **83**(14), 2934 (2003)
3. Ju, Y.S., Goodson, K.E.: Phonon scattering in silicon films with thickness of order 100 nm. *Appl. Phys. Lett.* **74**(20), 3005 (1999)
4. Song, D., Chen, G.: Thermal conductivity of periodic microporous silicon films. *Appl. Phys. Lett.* **84**(5), 687 (2004)
5. Maasilta, I., Minnich, A.J.: Heat under the microscope. *Phys. Today* **67**, 27–32 (2014)
6. Wilson, R.B., Cahill, D.G.: Anisotropic failure of Fourier theory in time-domain thermoreflectance experiments. *Nat. Commun.* **5**, 5075 (2014)
7. Chen, G.: Nanoscale Energy Transport and Conversion: A Parallel Treatment of Electrons, Molecules, Phonons, and Photons. Oxford University Press, New York (2005)
8. Esfarjani, K., Chen, G., Stokes, H.T.: Heat transport in silicon from first-principles calculations. *Phys. Rev. B* **84**, (2011). doi:10.1103/physrevb.84.085204
9. Ziman, J.M.: ELECTRON-ELECTRON INTERACTION, in *Electrons and Phonons*, pp. 159–174. Oxford University Press, Oxford (2001)
10. Davis, B.L., Hussein, M.I.: Nanophononic metamaterial: thermal conductivity reduction by local resonance. *Phys. Rev. Lett.* **112**, (2014). doi:10.1103/physrevlett.112.055505
11. Chen, G.: Multiscale simulation of phonon and electron thermal transport. *Annu. Rev. Heat Transf.* (2014). doi:10.1615/annualrevheattransfer.2014011051
12. Majumdar, A.: Microscale heat conduction in dielectric thin films. *J. Heat Transf.* **115**(1), 7 (1993)
13. Broido, D., Malorny, M., Birner, G., Mingo, N., Stewart, D.: *Appl. Phys. Lett.* **91**, 231922 (2007)
14. Romano, G., Grossman, J.C.: Multiscale phonon conduction in nanostructured materials predicted by bulk thermal conductivity accumulation function. arXiv preprint arXiv:1312.7849 (2013)
15. Minnich, A.J., Johnson, J., Schmidt, A., Esfarjani, K., Dresselhaus, M., Nelson, K.A., Chen, G.: Thermal conductivity spectroscopy technique to measure phonon mean free paths. *Phys. Rev. Lett.* **107**(9), 095901 (2011)
16. Regner, K.T., Sellan, D.P., Su, Z., Amon, C.H., McGaughey, A.J., Malen, J.A.: Broad-band phonon mean free path contributions to thermal conductivity measured using frequency domain thermoreflectance. *Nat. Commun.* **4**, 1640 (2013)
17. Chandrasekhar, S.: Radiative Transfer. Courier Dover Publications, New York (1960)
18. Nan, C.-W., Birninger, R., Clarke, D.R., Gleiter, H.: Effective thermal conductivity of particulate composites with interfacial thermal resistance. *J. Appl. Phys.* **81**(10), 6692 (1997)
19. Romano, G., Grossman, J.C.: Toward phonon-boundary engineering in nanoporous materials. *Appl. Phys. Lett.* **105**, 033116 (2014)
20. Klitsner, T., VanCleve, J., Fischer, H., Pohl, R.: Phonon radiative heat transfer and surface scattering. *Phys. Rev. B* **38**, 7576–7594 (1988)
21. Peterson, R.B.: Direct simulation of phonon-mediated heat transfer in a debye crystal. *J. Heat Transf.* **116**(4), 815 (1994)
22. Mazumder, S., Majumdar, A.: Monte Carlo study of phonon transport in solid thin films including dispersion and polarization. *J. Heat Transf.* **123**(4), 749 (2001)
23. Lacroix, D., Joulain, K., Lemonnier, D.: Monte Carlo transient phonon transport in silicon and germanium at nanoscales. *Phys. Rev. B* **72**, (2005). doi:10.1103/physrevb.72.064305
24. Hao, Q., Chen, G., Jeng, M.-S.: Frequency-dependent Monte Carlo simulations of phonon transport in two-dimensional porous silicon with aligned pores. *J. Appl. Phys.* **106**(11), 114321 (2009)
25. Peraud, J.-P. M., Hadjiconstantinou, N.G.: Efficient simulation of multidimensional phonon transport using energy-based variance-reduced Monte Carlo formulations. *Phys. Rev. B* **84**, (2011). doi:10.1103/physrevb.84.205331
26. Bird, G.: Molecular Gas Dynamics and the Direct Simulation of Gas Flows, vol. 1. Clarendon Press, Oxford (1994)
27. Homolle, T.M., Hadjiconstantinou, N.G.: A low-variance deviational simulation Monte Carlo for the Boltzmann equation. *J. Comput. Phys.* **226**, 2341–2358 (2007)
28. Radtke, G.A., Peraud, J.M., Hadjiconstantinou, N.G.: On efficient simulations of multiscale kinetic transport. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **371**, 20120182–20120182 (2012)
29. Peraud, J.-P.M., Hadjiconstantinou, N.G.: An alternative approach to efficient simulation of micro/nano-scale phonon transport. *Appl. Phys. Lett.* **101**(15), 153114 (2012)
30. Peraud, J.-P.M., Hadjiconstantinou, N.G.: Deviational phonons and thermal transport at the nanoscale. In: volume 7: Fluids and Heat Transfer Parts A, B, C, and D. ASME (2012)
31. Peraud, J.-P.M., Landon, C.D., Hadjiconstantinou, N. G.: Monte Carlo methods. *Annu. Rev. Heat Transf.* **17**, 205–265 (2014)
32. Peraud, J.-P.M., Hadjiconstantinou, N.G. (in press)
33. Hua, C., Minnich, A.J.: *Semicond. Sci. Technol.* **29**, 124004 (2014)
34. Ziman, J.M.: Electrons and Phonons. Clarendon, Oxford (1960)
35. Mingo, N., Stewart, D.A., Broido, D.A., Lindsay, L., Li, W.: Length-Scale Dependent Phonon Interactions, pp. 137–173. Springer, New York (2014)
36. Landon, C., Hadjiconstantinou, N.G.: *J. Appl. Phys.* **116**, 163502 (2014)
37. Wagner, G.J., Jones, R., Templeton, J., Parks, M.: *Comput. Methods Appl. Mech. Eng.* **197**, 3351 (2008)

Simulation of Supported Metal Clusters

Giovanni Barcaro¹, Luca Sementa² and Alessandro Fortunelli³

¹CNR-IPCF, Consiglio Nazionale delle Ricerche, Pisa, Tuscany, Italy

²CNR-ICCOM, Consiglio Nazionale delle Ricerche, Pisa, Tuscany, Italy

³Istituto per la Chimica dei Composti Organometallici (Institute for the Chemistry of Organometallic Compounds, ICCOM), Consiglio Nazionale delle Ricerche (National Research Council, CNR), Pisa, PI, Italy

producing enhanced magnetic moments and anisotropy, etc. All these properties are very sensitive not only to size, shape, and external fields but also – in the case of supported metal clusters – to the interaction of the cluster with the support. From the possibility of modifying the cluster properties by tuning cluster/substrate interactions, a great freedom in the physical and chemical behavior of these materials results and therefore their interest in science and technology. This interest has been translated into an intense theoretical effort aimed at the prediction and understanding of the structure and properties of metal clusters via simulations, which is the topic of the present entry.

Synonyms

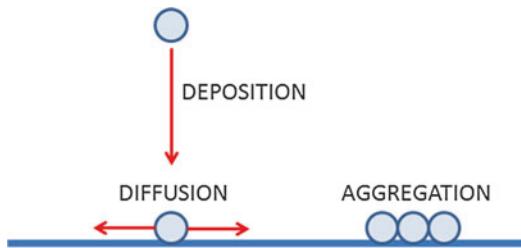
Deposited metal nanostructures; Heterogeneous catalysts; Nanoparticles; Nanoscale metals

Definition

A metal cluster is an aggregate composed of metallic elements, ranging in size between a few and few thousand atoms, while larger entities are usually referred to as “particles.” A metal cluster can exist in different forms: free, coated by a protective shell of ligands, or deposited on a support, in which the last case the expression “supported metal cluster” is employed. In these structures, nanoscale confinement effects appreciably modify the response properties of the metallic nanosystem [1, 2]. For example, the high surface/volume ratio assures that a great part of metal atoms lie at the surface in low-coordinated sites, which are therefore available for chemical sensing and catalysis. As another example, the collective oscillation of conduction band electrons which gives rise to plasmonic phenomena in bulk metals is modulated by the presence of metal/environment interfaces, thus producing the so-called surface plasmon resonances. Magnetic properties are also exalted by the reduction of the electronic density of states in such confined systems, hence

Introduction

A general problem in the application of metal clusters and nanoparticles is their intrinsic instability: due to their high surface/volume ratio, nanoparticles tend to decrease their energy by coalescing into larger particles (Ostwald ripening and sintering processes) [3]. Being intrinsically unstable, nanoclusters can only survive in the presence of kinetic barriers which avoid mass transfer and agglomeration processes, as due, for example, to specific interactions of nanoclusters with a support, with metal oxides as very common substrates [2]. Metal-on-oxide systems (as they are thus often called) can be created by depositing atoms or preformed clusters onto oxide surfaces in various heterogeneous environments (liquid/solid, gas/solid), followed by the processes of adsorption, diffusion, and self-organization of the aggregates (see Fig. 1). Supported metal clusters can be grown by chemical vapor deposition (CVD) or physical vapor deposition (PVD) techniques on top of a surface. In CVD the precursor metal atoms or clusters are protected by ligands which are eliminated via pyrolysis (the vaporized precursors are introduced into a reactor and adsorb onto a substrate held at elevated temperatures where they decompose producing metal species), whereas PVD involves condensation from the vapor phase of the bare precursor in the form of atoms or clusters. To control these processes



Simulation of Supported Metal Clusters,

Fig. 1 Surface processes in physical vapor deposition (PVD) of metal species onto a substrate and growth of nanoparticles

and the resulting structural, catalytic, optical, etc., properties, detailed information on metal/support interactions and elementary and collective mechanisms of growth and diffusion is needed.

The properties of metal clusters in fact often depend sensitively on the microscopic details of their structure. Among other tools, microscopy techniques such as high-resolution transmission electron microscopy (TEM) can be used to derive information on the geometric structure of a cluster with real-space resolution currently reaching the atomic scale. Especially interesting is atomic-resolution environmental TEM (ETEM) [4] which is able to operate under controlled and close-to-realistic conditions of pressure and temperature. An application of this technique is reported in Fig. 2, which shows the evolution of the shape of Cu clusters (supported on ZnO), under the action of a mixture of H₂, H₂O, and CO at different values of temperature and total pressure [5].

Supported metal clusters find important applications, most notably as heterogeneous catalysts [2]. In this respect metal clusters benefit not only from their high surface/volume ratio but also from the cluster greater structural freedom with respect to extended surfaces which can better accommodate incoming ligand species and promote their diffusion, with the presence of the underlying support also playing a role in tuning the cluster propensity, as pictorially exemplified in Fig. 3.

Such detailed geometric and mechanistic information as depicted in Figs. 1, 2, and 3 is difficult to obtain at the experimental level. Therefore, theory and simulation play a crucial

role in this field as complementary tools of characterization, analysis, and understanding. In this entry, a synthetic description of the approaches developed to meet this demand of theoretical information will be presented, especially focusing on the methods that are employed to predict the structural and electronic properties of supported metal clusters.

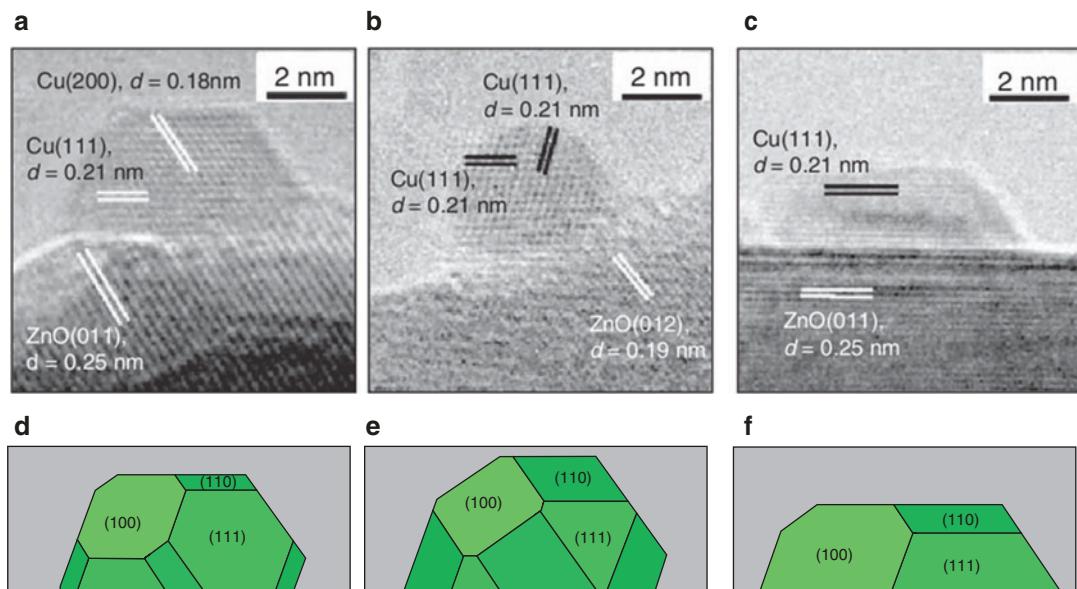
Simulation Methods

Calculations of energy and forces: to investigate the structural properties of supported metal clusters, what is needed is an expression for the system energy as a function of the structural degrees of freedom. In supported metal systems, especially for small clusters, a full atomistic description is usually employed, i.e., the degrees of freedom correspond to explicitly specifying the positions (the Cartesian coordinates) of all the atoms of the cluster and the support. One then needs to specify a function: E(X), where E is the energy and X are the atomic coordinates. Such an energy expression can be derived using different methods.

At the most fundamental level, one can use first-principles methods (i.e., methods not introducing parameters other than basic physical constants). Among these, the most widely employed in the field of metal clusters is the density functional theory (DFT), as it realizes the best compromise between chemical accuracy and computational effort. DFT in its most common form implies solutions of single-particle Kohn–Sham equations [7]:

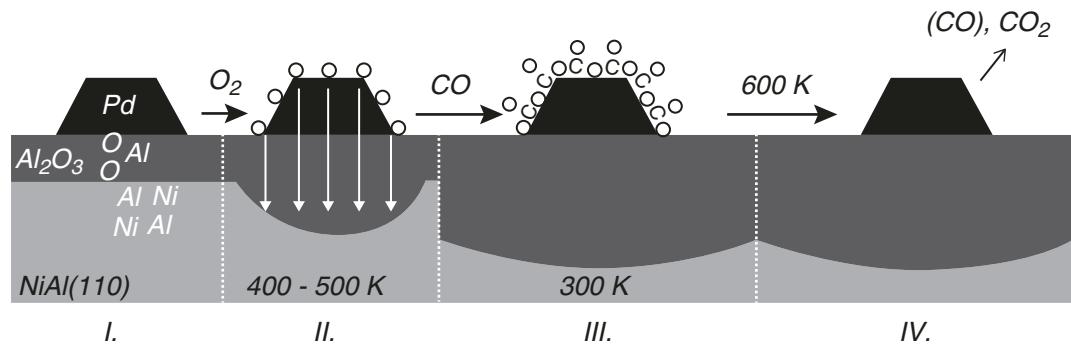
$$\left(-\frac{\hbar^2}{2m} \nabla^2 + V(\mathbf{r}, X) + \int \frac{(r')}{|r - r'|} dr' + V_{xc}[(\mathbf{r})](\mathbf{r}) \right) \varphi_i(\mathbf{r}) = \lambda_i \varphi_i(\mathbf{r}) \quad (1)$$

where the Hamiltonian operator (in parenthesis) is expressed as a functional of the electronic density $\rho(\mathbf{r})$ of the system and $\varphi_i(\mathbf{r})$ and λ_i are single-particle wave functions and energies, respectively. The Hamiltonian operator contains the contributions of kinetic energy, external potential



Simulation of Supported Metal Clusters, Fig. 2 In situ TEM images (a, b, and c) of a Cu/ZnO catalyst in various gas environments together with the corresponding Wulff constructions of the Cu nanocrystals (d, e, and f). (a) The image was recorded at a pressure of 1.5 mbar of H₂ at 220 °C. The electron beam is parallel to the [011] zone axis

of copper. (b) Obtained in a gas mixture of H₂ and H₂O, H₂:H₂O = 3:1 at a total pressure of 1.5 mbar at 220 °C. (c) Obtained in a gas mixture of H₂ (95 %) and CO (5 %) at a total pressure of 5 mbar at 220 °C (Reproduced with permission from Hansen et al. [5]. Copyright 2002 HighWire Press)



Simulation of Supported Metal Clusters, Fig. 3 The proposed mechanism for the interaction of Pd particles deposited on Al₂O₃/NiAl(110) with O₂ at elevated temperatures. It involves dissociation of oxygen on Pd, its diffusion through the oxide film, and reaction with the metallic substrate underneath (II). At temperatures above 400 K, the oxide film becomes thicker by about 3–5 oxygen

sub-layers. Surface oxygen reacts with CO during the reduction step (III), followed by CO and CO₂ desorption upon heating to 600 K (IV). The resulting surface exhibits adsorption–desorption properties very similar to the pristine samples (Reproduced with permission from Shaikhutdinov et al. [6]. Copyright 2002 Elsevier)

(depending on the coordinates X of the nuclei), coulombic repulsion, and exchange–correlation potential. DFT is a sophisticated and computationally demanding first-principles approach, with an explicit description of the electronic

wave function of the system, and can only be used in practice when the size of the system is relatively small – say, clusters between subnanometer and 2 nm as typical dimensions, even though advances in software and hardware

are continuously extending the scope of systems amenable to accurate DFT modeling.

When considering a supported cluster at the DFT level, one usually considers also an atomistic and first-principles description of the substrate. This gives rise to two separate issues.

First, it necessitates either employing a code with periodic boundary conditions (such that a unit cell encompassing the metal cluster and a piece of its underlying support is periodically replicated in two or three dimensions) or developing a finite-size model (usually named a “cluster” model, not to be confused with the metal cluster itself) which is embedded into the neighboring environment via properly chosen boundary conditions (such as an external potential or pseudo-atoms mimicking the environment). The use of DFT periodic computational codes is currently the most common choice. The explicit description of the substrate entails a remarkable increase of the computational cost, thus resulting in a limitation in the size of the systems which can be treated at this level of theory.

Second, the substrate itself may not be properly described by a standard DFT approach. Considering, for example, transition metal or rare earth oxide supports, one can face the problem that standard DFT tends to excessively delocalize the *d*-band or *f*-band electrons and simultaneously decrease the associated magnetism. A convenient way to circumvent the problem is to add an energy term to the Hamiltonian operator: the Hubbard *U* term, so named after its proposer [8], which corresponds to an energy penalty for doubly occupying a given orbital. The *U* term can be self-consistently inserted into the Kohn–Sham operator of Eq. 1, and by choosing a proper value for the *U* term, one can achieve a much better description of the oxide support. The ensuing method is called DFT + *U*.

For larger metal aggregates, say over 2 nm in size, DFT becomes excessively demanding, and one has to switch to a description of the system in terms of analytic potentials. In these approaches the energy function $E(\mathbf{X})$ is expressed in terms of analytic expressions of the atomic coordinates. As the number of such coordinates is huge, further simplifications must be sought. Let us focus first

on the metal–metal interactions. A common practice is to first define $E(\mathbf{X})$ as a sum of site atomic energies $E^{i,\text{atomic}}(X)$:

$$E = \sum_i E^{i,\text{atomic}}(X) \quad (2)$$

where the sum runs over all *i*-atoms. One then expresses the dependence of the $E^{i,\text{atomic}}(X)$ functions on the coordinates of atoms other than the *i*th through generalized collective variables (GCV). The GCV, i.e., are a limited number of functions which describe the distribution of neighbors around the given atom and depend on an analytically compact form upon their number, distance, and orientation. What distinguishes the various analytic potentials is the choice of the GCV and the analytic form by which $E^{i,\text{atomic}}$ depend on the GCV. One of the simplest analytic potentials to describe the metal–metal interactions is the second moment approximation (SMA) tight-binding potential [9], in which each atomic energy is written as a sum of a pair repulsive ($E^{i,\text{rep}}$) and a many-body attractive ($E^{i,\text{att}}$) component:

$$E = \sum_{i=1}^N \{E^{i,\text{rep}}(i) + E^{i,\text{att}}(i)\} \quad (3)$$

where

$$E^{i,\text{rep}}(i) = \sum_{j \neq i}^N A \exp \left\{ -p \left(\frac{r_{ij}}{r_0} - 1 \right) \right\} \quad (4)$$

$$E^{i,\text{att}}(i) = \left[\sum_{j \neq i}^N \xi^2 \exp \left\{ -2q \left(\frac{r_{ij}}{r_0} \right) - 1 \right\} \right]^{1/2} \quad (5)$$

and the sums in Eqs. 3, 4, and, 5 run over all *j*-atoms different from the *i*th. The parameters *A*, r_0 , ξ , p , and q are characteristic of the given metal element and are usually fitted to the experimental quantities such as the cohesive energy, the equilibrium lattice parameter, and independent elastic constants for the reference crystal structure. It should be stressed that in the SMA analytic potential, the repulsive term in $E^{i,\text{atomic}}(X)$ is linear, i.e.,

simply additive, whereas the attractive term has a many-body character and is highly nonlinear since it is proportional to the square root of a sum of additive terms. This *many-body* or *nonlinear behavior* as a function of coordination is an essential feature of the metallic bond. This feature allows metallic elements to sustain a great number of possible coordination numbers whose strength however tends to *saturate* as the coordination number increases.

While many different variants of analytical potentials are available for describing metal–metal binding, the choice is much more restricted for the metal/oxide interaction. One of the most accurate so far proposed refers to the square-symmetry MgO(100) surface and is one in which the substrate is assumed to be rigid, i.e., the oxide atoms are assumed to be frozen in their equilibrium configurations [10]. The metal/oxide interaction is then expressed as a sum of the contribution coming from each metal atom:

$$E^{Pd-MgO} = \sum_{i=1}^N E_i(x, y, z, C) \quad (6)$$

where N is the number of metal atoms in the cluster. The atomic energies E_i depend upon the spatial position (x, y, z) and the metal coordination (C) of each metal atom. The dependence upon the height of the metal atom with respect to the substrate plane (x, y) is described via a simple Morse-type function:

$$E_i(x, y, z, C) = a_1 + e^{-2a_2(z-a_3)} - 2e^{-a_2(z-a_3)} \quad (7)$$

where the parameters a_i , $i = 1–3$, depend on (x, y) and the coordination number C . It should be noted that the dependence of the a_i parameters on metal coordination is *nonlinear* or *many body*. In fact, it is intuitive that the metal/oxide interaction is strongest in the case of adsorption of isolated adatoms and decreases progressively in a nonlinear way as a function of an increasing number of metallic neighbors. A three-point interpolation formula describing an exponentially decreasing dependence of the interaction energy

on the coordination of the metal atom C is used in the parametrization which hence reads

$$a_i(x, y, C) = b_{i,1} + b_{i,2}e^{-C/b_{i,3}} \quad (8)$$

The parameters $b_{i,j}$, $i, j = 1–3$, finally depend on (x, y) via a linear combination of trigonometric functions preserving the square symmetry of the MgO(100) unit cell:

$$\begin{aligned} b_{i,j}(x, y) = & c_{i,j,1} + c_{i,j,2}[\cos(x) + \cos(y)] \\ & + c_{i,j,3}[\cos(x+y) + \cos(x-y)] \end{aligned} \quad (9)$$

In this way, a set of $27 c_{ijk}$, $i, j, k = 1–3$, parameters is sufficient to describe the oxide/metal interaction. This procedure has been extended and successfully applied to other metals interacting with the same oxide surface. Unfortunately, analytic potentials describing the metal/oxide interaction in the case of nonrigid oxide substrates are not yet available.

Algorithms for structural search: supported metal clusters can exhibit an impressive variety of different structures and morphologies (see Ref. [11] and the discussion below). In the case of small clusters (below 100–200 metal atoms), the majority of the atoms are on the surface, whence the importance of reducing the surface energy. This reduction can be achieved by adopting noncrystalline structures, which manage to decrease the surface energy at the expense of increasing the cluster internal energy due to bond strain. The Mackay icosahedron is probably the most common noncrystalline motif: this structure is limited by (111) close-packed facets only, which are usually the surfaces with the lowest energy, thus minimizing the cluster surface energy [11]. However, Mackay icosahedra are highly stressed structures (radial inter-shell bonds are highly compressed, whereas intra-shell bonds are highly expanded) and are hence favorable only at small sizes. Moreover, their interaction with a substrate is often disfavored so that they are not common in supported metal clusters. Another noncrystalline motif is represented by the decahedron: this structure is formed by two pentagonal

pyramids sharing a base; its surface has only (111) close-packed facets. A truncation of the five edges limiting the common basis of the pyramids to expose (100)-like facets is advantageous and produces the so-called Ino decahedron. Other cuts with reentrances exposing further (111) close-packed facets separating neighboring (100)-like facets can also be advantageous (Marks decahedra). Decahedra are also strained structures, but the strain is much smaller than for icosahedra, making them favorable in an intermediate range of sizes [11]. When the cluster grows in size, crystalline motifs become preferred as the ratio between surface and bulk atoms quickly decreases: volume contribution is hence optimized, at the expense of a higher surface energy.

All these motifs can be adopted by both isolated and supported metal clusters. Moreover, such motifs are also often close in energy, so that the metal clusters can transition among them, i.e., exhibit what is usually called a “fluxional” behavior. The fact that the PES (potential energy surface) of a metal cluster system is characterized by a large number of local minima with relatively small energy differences among them represents a challenge at the theoretical/computational level if accurate quantitative predictions are sought for. To be predictive, it is necessary to utilize theoretical methods able to perform a systematic sampling of the system PES. Appropriate methods to this effect are based on global optimization techniques, i.e., computational techniques aimed at finding the absolute minimum (or global minimum) of a mathematical function which can be very complex and exhibit very many minima differing by small energy amounts.

One of the simplest and most efficient global optimization techniques is the so-called Monte Carlo with minimization or basin-hopping (BH) algorithm, which has been proposed more than 25 years ago [12] but is still the most commonly used. The BH algorithm consists of the following steps:

1. An initial random configuration of the metal cluster is chosen, a local geometry optimization is performed, and the final energy (the

so-called “fitness” parameter) is registered as E_1 .

2. Starting from the relaxed configuration, the atoms of the metal cluster are subjected to a random move, a new local geometry optimization is performed, and the final energy is registered as E_2 .
3. A random number (rndm) between 0 and 1 is generated, and the movement of step 2 is accepted only if $\exp[(E_1 - E_2)/k_B T] > \text{rndm}$ (Metropolis criterion).
4. Steps 2 and 3 – the Monte Carlo steps – are repeated a given number of times.

Depending on the $k_B T$ parameter, which plays the role of a fictitious temperature, some high-energy configurations are accepted, and the search is able to explore different structural motifs (belonging to different funnels of the PES) of the metal cluster. The thoroughness of the BH search is much increased when the BH protocol is coupled with *structural recognition* algorithms, i.e., algorithms which classify geometrical configurations as belonging to a given structural motif (or structural family) [11].

The BH algorithm can be in principle combined with any method for calculating energy and forces, but the feasibility of such a systematic sampling approach depends on the size of the system, so that a multi-scale hierarchy of theoretical methods can be envisaged [11].

For small supported metal clusters (with a number of atoms, N , less than a few tens, say $N \leq 40$), it is computationally feasible to conduct a global optimization (in the form of the BH algorithm) using energy and forces derived from DFT in a DF-BH method.

For medium-size supported metal clusters (say, $40 \leq N \leq 200$), the DF-BH approach becomes progressively less feasible, and more approximate methods such as analytic potentials (AP) must be employed. An efficient possibility is to replace the DF-BH approach with a combination of DFT and AP simulations, in the DF-AP method. In this approach an AP for the given system is first selected (or developed). Second, a thorough global optimization search using, e.g., the BH technique with structural recognition algorithms

is conducted using this AP. Third, the low-energy configurations produced by the BH search are grouped into structural families or basins. Fourth, a few of the low-energy configurations belonging to each structural basin are subjected to local geometry relaxation using a DFT method, and the changes in the relative energy ordering of the structural families are analyzed. Finally, if these changes highlight inconsistencies or inaccuracies in the AP, the DFT results are used to refine the AP in a self-consistent process.

For yet larger supported metal clusters (say $200 \leq N \leq \text{few} \cdot 1,000$), even performing DFT calculations becomes progressively unfeasible, at least conducting them in large numbers, and one has to rely entirely on global optimization searches based on AP approaches. The size of the system justifies the use of an AP and thus an average description of the metallic bond.

At some point (the current limit is around 1,000–2,000 supported metal atoms) even the BH algorithm using AP becomes computationally unaffordable. An efficient alternative solution is to use extrapolations based on structural motifs, which is discussed in the following subsection. The idea is that for these large particles the competing motifs have been singled out and are known, so that large clusters belonging to these motifs can be constructed and locally optimized and their energy can be compared to predict energetic crossover and phase transformations among structural families.

Large Particles: Generalized Wulff Construction

Let us assume that the structural family of the metal cluster is known, e.g., it corresponds to a crystalline motif. In this family, the atoms of the clusters are positioned in the crystalline sites characterizing the metal in its bulk form (e.g., face centered cubic, fcc). Still, the cluster shape needs to be determined. The “best” way to cut a gas-phase (or isolated) fcc particle was investigated by Wulff [13], and the corresponding building criterion (originally proposed for macroscopic aggregates but which can be translated into

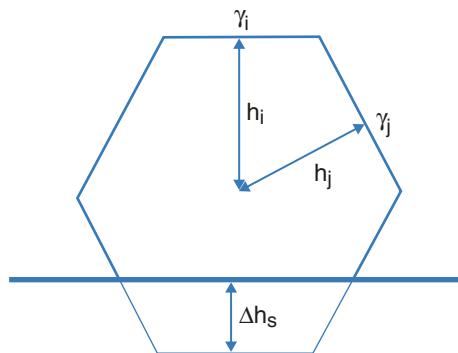
atomistic systems) goes under the name of Wulff construction. In this construction, taking for simplicity a crystal only limited by (111) and (100) faces, the best shape fulfills the following condition:

$$\frac{\gamma(100)}{\gamma(111)} = \frac{d(100)}{d(111)} \quad (10)$$

where $\gamma(100)$ and $\gamma(111)$ are the surface energies of the (100) and (111) faces, respectively, whereas $d(100)$ and $d(111)$ are the distances of the facets from the center of the cluster. The Wulff criterion has a general validity and can also be applied to higher-Miller-index surfaces – as the (110), etc. – and also noncrystalline structures, once low-energy surfaces are determined in these motifs.

The Wulff construction is valid for isolated clusters or particles. When considering the equilibrium structure of a cluster which is in contact with an environment, the Wulff construction can be straightforwardly generalized by considering the role played by the environment. Suppose, for example, that the cluster is surrounded by an atmosphere of interacting molecules. The adsorption of such species on the facets of the cluster will modify the corresponding surface energies. A generalized Wulff construction in which the *bare* surface energies are replaced by surface energies corresponding to facets “vested” by the ligand species can be used. This often gives rise to a phenomenon known as *particle restructuring* or *reshaping* and leads to substantial changes in the particle morphology, as observed experimentally in fair agreement with theoretical predictions (see, e.g., Ref. [14]). For example, in reactive oxidative conditions one typically expects that more open facets such as (100) are more reactive with respect to (111) and therefore are more stabilized by the interaction with oxygen species, so that they will grow at the expense of the more inert ones with respect to more reducing conditions.

A similar, properly modulated reasoning holds for supported clusters, and the corresponding generalized Wulff construction goes under the name of Wulff–Kaishev construction [15]. Suppose for the moment to ignore the influence of adsorbed



Simulation of Supported Metal Clusters,

Fig. 4 Schematic representation of the equilibrium shape of a supported crystal. The Wulff shape of the free crystal is truncated at the interface by Δ_{hs} , which is proportional to the adhesion energy

species. A supported cluster can then be considered as isolated except for one face which is in contact with the substrate. It suffices to change the surface energy of such facet by taking into account the adhesion energy to the substrate to obtain the Wulff–Kaishev construction, illustrated in Fig. 4.

As it can be observed in Fig. 4, starting from the equilibrium shape dictated by the Wulff construction, a portion of the cluster can be cut off more or less profoundly depending on the adhesion energy, thus resulting in an asymmetric shape with respect to the central plane of the particle. In other words, the adhesion energy of the particle to the substrate reduces the surface energy of the particle face in contact with the substrate, also reducing the distance between this surface and the center of the nanoparticle. A simple example taking into account both the influence of ligand species and a support is given in Fig. 5, where reshaping of Ag clusters deposited on an amorphous alumina support under an O₂ atmosphere is depicted [14].

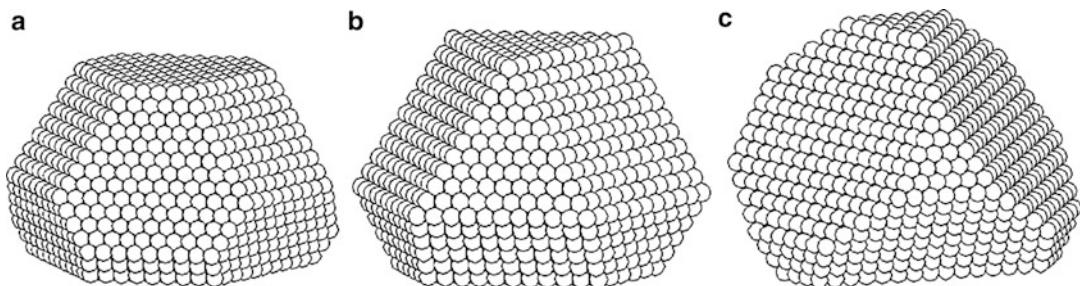
From the discussion of the Wulff–Kaishev construction, it results that it is crucial to be able to predict accurately the adhesion energy of a cluster to its support. In general, the adhesion of a metal cluster to a substrate depends on the detailed atomistic relationships between the metal particle and the substrate, also called

epitaxy or epitaxial relationships. As an example, let us examine the results of global optimization searches carried out on four pure metals (Ag, Au, Pd, and Pt) adsorbed on MgO(100) [16]. The simulations on Ag, Au, Pd, and Pt particles adsorbed on MgO(100) singled out three dominant motifs (see Fig. 6):

1. Fcc clusters in (001) epitaxy
2. Fcc clusters in (111) epitaxy
3. Decahedral cluster which adhere to the substrate with a pseudo-(001) facet

Icosahedral clusters do not achieve a good matching with the substrate and are therefore not favorable for these metal/support combinations. Decahedra can lie on a pseudo-(100) facet and achieve a better matching. Concerning fcc clusters, the competition between the (001) and the (111) epitaxy is due to the following reason. The (001) epitaxy has a stronger adhesion energy per unit contact area, due to its better matching with the oxygen atoms of the square-symmetry substrate, especially when the lattice mismatch between metal and oxide is small. However, in the (111) epitaxy this can be compensated by a larger contact area, so that the total adhesion energy may become even larger than in the (001) epitaxy.

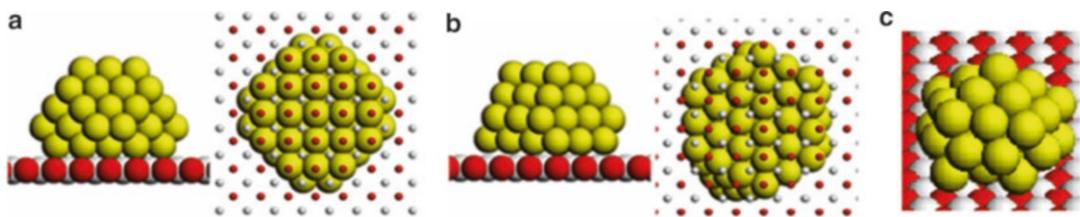
The previous example shows how subtle can be the competition among different epitaxies in supported clusters. This becomes even more complicated when the substrate does not simply play a weak “perturbative” effect on the metallic particle equilibrium shape, but a strong interaction between the metallic cluster and the substrate completely alters the equilibrium shape of the former with respect to the gas phase. The simplest way in which this reconstruction can be realized is via the creation of a network of dislocations. This solution is often found in larger aggregates and thin films, for which a substantial adhesion energy drives the system to better match the substrate but is not sufficient to compensate the loss of metallic energy within the particle due to a transition to a completely different structural motif. An alternative possibility is that the interaction with the support can drive the metal cluster toward exotic



Simulation of Supported Metal Clusters,

Fig. 5 Morphology of amorphous-alumina-supported silver particles in case of (a) nonoxidized surfaces and oxygen-covered surfaces at (b) $p(O_2) = 5 \times 10^{-3}$ atm

and (c) $p(O_2) = 1$ atm. The lateral dimension of the clusters is about 6 nm (Reproduced with permission from Molina et al. [14]. Copyright 2011 Elsevier)



Simulation of Supported Metal Clusters, Fig. 6 Metal clusters deposited on a $MgO(100)$ surface. (a) Side and bottom views of a fcc (001) cluster. (b) Side and bottom views of a fcc (111) cluster. (c) Side/top view of a decahedral cluster. Oxygen atoms are in red and

magnesium atoms are in light gray. In the bottom views (right part of a and b), substrate atoms are represented by small spheres so that the contact epitaxy of the cluster is visible (Reproduced with permission from Ferrando et al. [16]. Copyright 2009 American Institute of Physics)

interface-stabilized structures which may even have no counterpart in the gas phase. A well-understood example is given by Ni clusters grown on $MgO(100)$ [17]. In this system Ni atoms would like to be positioned on top of oxygen atoms of the support but – due to the large size mismatch between the nearest-neighbor distance in bulk Ni and the O–O distance in the MgO substrate – the preferred structural motif turns out to be hcp-like (note that Ni is fcc in the bulk), with close-packed planes alternating in ...ABAB... stacking and oriented perpendicular to the oxide substrate along its [100] direction (see Fig. 7). Theoretical simulations show that, as the size of the clusters increases above $N = 40$, the lowest-energy structures are hcp for almost all sizes until for clusters larger than about 2,500 atoms, the interfacial energy is not sufficient anymore to compensate the transformation from fcc to hcp structure, and fcc-like clusters with a

(111)/(100) epitaxial relationship to the support become energetically favorable. These predictions are in excellent agreement with experiment [17].

Electronic Effects

To conclude this entry, electronic effects, of great importance in catalysis, are briefly discussed.

The cluster electronic structure can be significantly affected by the interaction with the support: phenomena such as the quenching or at least lowering of the spin state of the metal cluster upon adsorption or charge transfer effects at the support/cluster interface are well documented [2, 11].

In general, when describing the metal/oxide interaction, it is customary to single out four main components: charge transfer, covalent binding, polarization, and dispersion interactions.

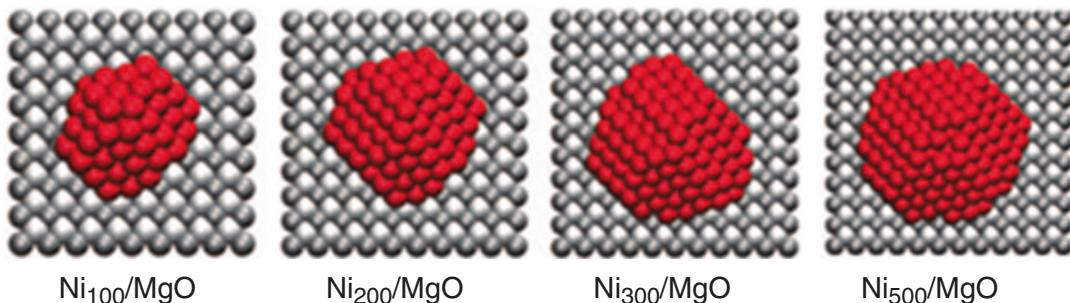
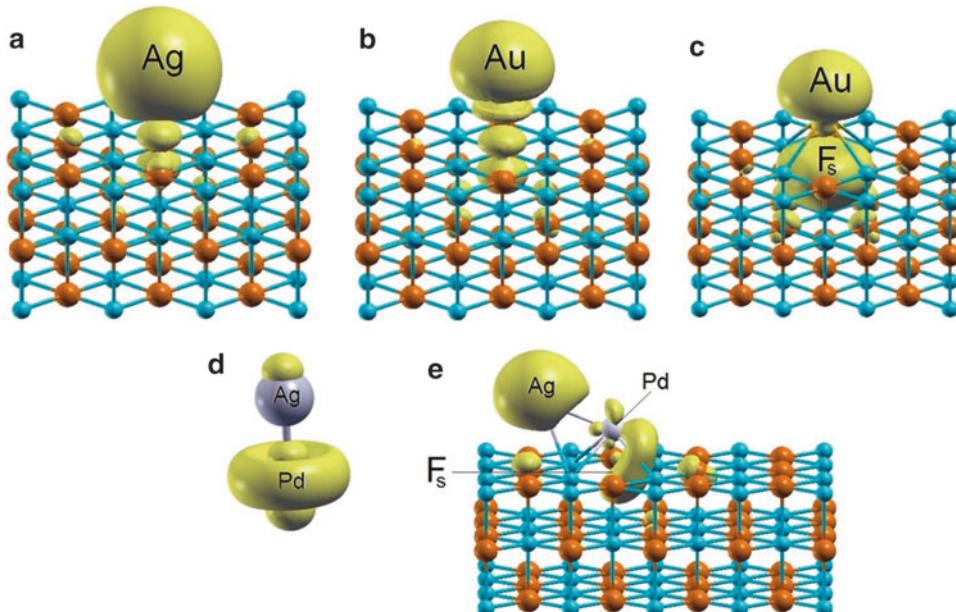

Simulation of Supported Metal Clusters,

Fig. 7 Global minima of Ni/MgO nanodots (of hcp structure) shown for $N = 100$, 200, 300, and 500 atoms

(Reproduced with permission and adapted from Ferrando et al. [17]. Copyright 2008 American Chemical Society)



Simulation of Supported Metal Clusters, Fig. 8 Spin density plots of (a) an Ag and (b) an Au adatom interacting with the regular (100) MgO surface, (c) an Au adatom interacting with the F_s -defected (100) MgO surface, (d)

the gas-phase PdAg dimer, (e) the PdAg dimer interacting with the F_s -defected (100) MgO surface. Isosurfaces at a value of 0.001 AU are plotted (Reproduced with permission from Negreiros et al. [19]. Copyright 2014 Elsevier)

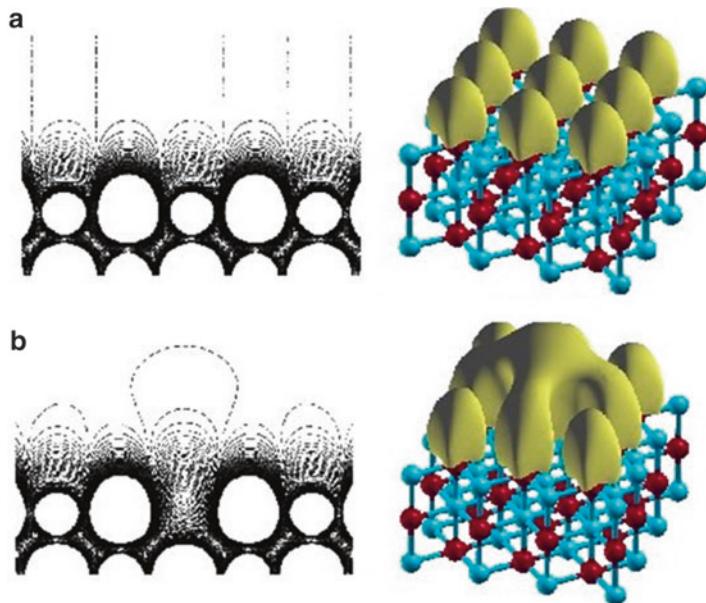
On simple oxides, it is often found that charge transfer between the metal and the oxide atoms is not large on regular surfaces. It can however be increased when the work function of the oxide is reduced by creating electronic states in the band gap via defects such as an oxygen vacancy. Charge transfer from/to the metal cluster can also be appreciable on oxides which act as a medium for electron transfer from an underlying

electron reservoir as it occurs for oxide ultrathin films grown on low-work-function metal surfaces [18].

As an illustration, Fig. 8 shows a comparison of plots of the spin density for different systems. From this figure the changes due to a change in metal and the present of defects are apparent: Ag on the regular $\text{MgO}(100)$ surface exhibits weak charge transfer, Au on the regular surface exhibits

Simulation of Supported Metal Clusters,

Fig. 9 Electrostatic potential generated by (a) the regular and (b) an F_s -defected (100) MgO surface is shown in a 2D map (left side) and 3D map (right side). In the case of the 3D map, an isosurface of 0.0005 AU is plotted (Reproduced with permission from Barcaro et al. [20]. Copyright 2009 Springer)



a less weak charge transfer (and correspondingly a bit stronger bonding), and Au on an F_s center exhibits strong bonding and a very strong charge transfer. In Fig. 8 the case of a binary dimer (AgPd) is also shown: one can see that AgPd in the gas phase contains a Pd atom whose electronic configuration has been promoted from d^{10} (which is the configuration of the free atom) to d^9s^1 (which corresponds to the valence state of the fully formed metallic bond in bulk systems), whereas when AgPd is deposited on an F_s center of MgO(100), the electronic configuration of Pd is switched back from d^9s^1 to d^{10} ; this shows that the interaction with the substrate can decrease magnetism (of Pd in this case) as mentioned above. Note also that, as the chosen metal systems contain an unpaired electron, the spin density gives good indications on chemical bond effects.

In addition to charge transfer, one less recognized effect of an oxide support is the strong polarization due to the electric field generated by such a charge-separated substrate. Even on the regular MgO(100) surface (see Fig. 9 for a plot of the electrostatic potential generated by the regular oxide surface, in which the electrostatic potential generated by an F_s -defected surface is also shown for comparison) and thus for

metal/oxide interactions without or with little charge transfer, this electric field can completely alter the PES of small supported clusters. For example, breaking of O_2 on small Au_3 and Ag_3 clusters becomes thermodynamically favorable, opposite to the situation in the gas phase [11]. The effect of the support electrostatic field is connected with the polarization component of the metal/oxide interaction and can be important for both small and larger clusters. This effect is also the origin of the so-called “metal-on-top” effect, i.e., the enhancement of metal adhesion to the substrate due to the presence of other metal atoms on top of those which are directly in contact with the substrate [11]. The metal-on-top effect can play an important role in determining the structure of supported metal clusters, e.g., favoring upright with respect to epitaxial configurations of small metal clusters.

Conclusions

In conclusion, simulations represent an effective tool to characterize and predict the structure and properties of supported metal clusters, i.e., small- or intermediate-size metal aggregates deposited

on a support. Several theoretical approaches are used in this field, ranging from first-principles methods to analytic potentials to simplified constructions based on surface and adhesion energetics. In conjunction with these methods, global optimization explorations of the potential energy surfaces of such systems allow one to accurately predict their structural features, which serve as a basis for a deeper understanding and knowledge of their striking structural, catalytic, optical, magnetic, etc. response properties. Electronic interactions between the cluster and its support are particularly interesting, as they affect profoundly the activity and efficiency of supported clusters in heterogeneous catalysis, one of their main technological applications. The substrate can play a role simply as a stabilizing or anchoring support or as a component actively interacting with and strongly modifying the properties of the metal cluster, and this two-fold role has been underlined here. Attention has been mostly focused on general phenomena, while more detailed information on properties other than structural (and the theoretical approaches used to predict them) can be found in the cross-reference essays.

Cross-References

- [Ab initio DFT Simulations of Nanostructures](#)
- [Bond-Order Potential](#)
- [Cluster](#)
- [Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling](#)
- [Engineered Nanoparticles](#)
- [Integrated Approach for the Rational Design of Nanoparticles](#)
- [Magnetic Nanostructures and Spintronics](#)
- [Metal Nanoparticles from First Principles](#)
- [Multiscale Modeling](#)
- [Nanoalloy Simulation](#)
- [Nanocluster](#)
- [Nanoscale Particle](#)
- [Optimization of Nanoparticles](#)
- [Mechanical Properties of Nanocrystalline Metals](#)

References

1. McWeeny, R.: Coulson's Valence, 3rd edn. Oxford University Press, Oxford (1980)
2. Ertl, G., Knotziger, H., Schuth, F., Weitkamp, J. (eds.): Handbook of Heterogeneous Catalysis, 2nd edn. Wiley, New York (2008)
3. Parker, S.C., Campbell, C.T.: Reactivity and sintering kinetics of Au/TiO₂(110) model catalysts: particle size effects. *Top. Catal.* **44**, 3–13 (2007)
4. Giorgio, S., Joao, S.S., Nitsche, S., Chaudanson, D., Sitja, G., Henry, C.R.: Environmental electron microscopy (ETEM) for catalysts with a closed E-cell with carbon windows. *Ultramicroscopy* **106**, 503–507 (2006)
5. Hansen, P.L., Wagner, J.B., Helveg, S., Rostrup-Nielsen, J.R., Clausen, B.S., Topsoe, H.: Atom-resolved imaging of dynamic shape changes in supported copper nanocrystals. *Science* **295**, 2053–2055 (2002)
6. Shaikhutdinov, S., Heemeier, M., Hoffmann, J., Meusel, I., Richter, B., Bäumer, M., Kuhlenbeck, H., Libuda, J., Freund, H.J., Oldman, R., Jackson, S.D., Konvicka, C., Schmid, M., Varga, P.: Interaction of oxygen with palladium deposited on a thin alumina film. *Surf. Sci.* **501**, 270–281 (2002)
7. Kohn, W., Sham, L.J.: Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, 1133A (1965)
8. Hubbard, P.S.: *Rev. Mod. Phys.* **33**, 249–264 (1961)
9. Cleri, F., Rosato, V.: Tight-binding potentials for transition metals and alloys. *Phys. Rev. B* **48**, 22–33 (1993)
10. Vervisch, W., Mottet, C., Goniakowski, J.: Theoretical study of the atomic structure of Pd nanoclusters deposited on a MgO(100) surface. *Phys. Rev. B* **65**, 245411 (2002)
11. Fortunelli, A., Barcaro, G.: Density-functional theory of free and supported metal nanoclusters and nanoalloys. In: Mariscal, M.M., Oviedo, O.A., Leiva, E.P.M. (eds.) *Metal Clusters and Nanoalloys: From Modeling to Applications*. Springer, Berlin (2013). doi:10.1007/978-1-4614-3643-0_2
12. Li, Z., Scheraga, H.A.: Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **84**, 6611–6615 (1987)
13. Wulff, G.: On the question of speed of growth and dissolution of crystal surfaces. *Z. Krystallogr.* **34**, 449–530 (1901)
14. Molina, L.M., Lee, S., Sell, K., Barcaro, G., Fortunelli, A., Lee, B., Seifert, S., Winans, R.E., Elam, J.W., Pellin, M.J., Barke, I., von Oeynhausen, V., Lei, Y., Meyer, R.J., Alonso, J.A., Rodriguez, A.F., Kleibert, A., Giorgio, S., Henry, C.R., Meiws-Broer, K.H., Vajda, S.: Size-dependent selectivity and activity of silver nanoclusters in the partial oxidation of propylene to propylene oxide and acrolein: a joint experimental and theoretical study. *Catal. Today* **160**, 116–130 (2011)

15. Kaischew, R.: Arbeitstagung Festkorper Physik, Dresden, p. 81 (1952)
16. Ferrando, R., Rossi, G., Levi, A.C., Kuntová, Z., Nita, F., Barcaro, G., Fortunelli, A., Jelea, A., Mottet, C., Goniakowski, J.: Structures of metal nanoparticles adsorbed on MgO(001). I. Ag and Au. *J. Chem. Phys.* **130**, 174702 (2009)
17. Ferrando, R., Rossi, G., Nita, F., Barcaro, G., Fortunelli, A.: Interface-stabilized phases of metal-on-oxide nanodots. *ACS Nano* **2**, 1849–1856 (2008)
18. Repp, J., Meyer, G., Olsson, F.E., Persson, M.: Controlling the charge state of individual gold atoms. *Science* **305**, 493–495 (2004)
19. Negreiros, F., Barcaro, G., Sementa, L., Fortunelli, A.: Concepts in theoretical heterogeneous ultranano-catalysis. *C. R. Chim.* **17**, 625–633 (2014)
20. Barcaro, G., Causà, M., Fortunelli, A.: A comparison between the adsorption properties of the regular and F_s-defected MgO(100) surface. *Theor. Chem. Acc.* **118**, 807–812 (2007)

Simulations of Bulk Nanostructured Solids

Donald W. Brenner

Department of Materials Science and
Engineering, North Carolina State University,
Raleigh, NC, USA

Synonyms

Plasticity of nanostructured solids

Definition

A bulk nanostructured solid is a material available in macroscopic quantities that has nanometer-scale structural features in at least one dimension.

Computer modeling has emerged to become a third pillar of research alongside experiment and theory. There are many reasons for this emergence, including an exponential increase in processing speed, relatively inexpensive platforms for parallel computing and data storage, new visualization capabilities, and the development of powerful algorithms that take full advantage of these advances in hardware. Advances in simulation methodologies have also made the

results of atomic-level computer modeling reliable to the extent that in many cases simulations can replace difficult and expensive experiments. Methods that enable atomistic dynamics using full electronic energies now allow processes involving up to several thousand atoms to be accurately modeled, with significantly larger systems on the horizon. Similarly, advances in materials theory have produced relatively simple analytic potential energy functions that can capture essentials of quantum mechanical bonding for simulations involving well over a billion atoms [1].

Computer modeling has played an especially important role in developing the current understanding of nanometer-scale structures and processes [2]. Indeed, among the many scientific and technological advances coming from nanotechnology is the ability for computer modeling and experiment to characterize phenomena on a common scale. Atomic-level computer modeling is now commonly used to explore and even predict new properties and processes that can be probed experimentally, to suggest new materials and structures with unique and desirable properties, to provide insight into the results of experiments, to generate data for larger-scale analysis, and to test scaling laws and analytic theories. In the case of nanostructured solids, computer modeling is allowing researchers to understand at the atomic level the structure, deformation mechanisms, and thermal-mechanical properties of these new materials with unprecedented detail [3, 4].

The two standard methods for modeling nanometer-scale systems are molecular dynamics and Monte Carlo simulation. In the former, coupled classical equations of motion for the atoms are integrated stepwise in time. Time steps generally range from one-half to tens of femtoseconds depending on the highest frequency vibration of interest, and simulations are typically carried out for picoseconds to tens of nanoseconds depending on system size, the phenomena to be studied, and the method in which interatomic forces are calculated. In a typical equilibrium Monte Carlo simulation, atomic configurations are generated with a probability that is proportional to their Boltzmann factor. Thermodynamic quantities in a Canonical ensemble are then obtained by averaging over the

properties of each configuration. Alternatively, in kinetic Monte Carlo simulation, time-ordered configurations are generated, typically using some rate expression.

Equilibrium Monte Carlo modeling requires specifying a potential energy as a function of atomic positions to calculate Boltzmann factors, while molecular dynamics simulations require interatomic forces, which are typically obtained as partial derivatives of the potential energy. In general, two approaches are used to calculate interatomic energies and forces. In the least computationally intensive approach, the interactions between atoms due to the electrons are replaced with effective interactions that are described with an analytic potential energy function. At present, there is no definitive mathematical form for the potential energy function, and forms that range from relatively simple pair-additive expressions to complicated many-body forms are used depending on the system, type of bonding, and phenomena to be studied. In the second approach to calculating potential energies, explicit electronic degrees of freedom are retained, the energy for which is calculated either using first principles methods or through a simplified semiempirical Hamiltonian. The calculation of total energies from first principles is well defined in terms of basis sets, electron correlation for ab initio methods, or choice of density functional expression (and pseudopotential) for density functional theory calculations. This is in contrast to analytic potentials, where a set of parameters (and often entirely new functional forms) must be developed for each system.

Nanostructured materials can be defined as materials that have at least one dimension in the “nanoscale” (typically 1–100 nm). Depending on which dimension this is, they can be classified into nanoparticles, layered or lamellar structures, filamentary structures, and bulk nanostructured materials. This entry is focused on the latter, which can be thought of as a traditional material with grain sizes at the nanometer scale. The nanometer-scale grains introduce unusual properties compared to more traditional micron-scale grain sizes, including unique combinations of strength and ductility as discussed in more detail below. The origin of these differences, and how they are being revealed

by computer modeling, makes up the bulk of the material discussed below.

Atomic positions for bulk nanostructured solids have been generated by a number of different methods. Some of the methods are based on geometrical constraints imposed by the simulation conditions, for example, ensuring that active slip systems are properly oriented with respect to periodic boundaries [5, 6]. Crystallization dynamics have also been used to generate nanostructures. These methods can be based on a Johnson-Mehl or a Potts model construction, both of which produces a log normal grain size distribution, or by using a molecular dynamics simulation to model crystallization from a melt [3, 7, 8]. Other researchers have generated nanostructures by simulating compaction and sintering of nanoclusters [9–11]. Other methods use random grain centers, with grain boundaries chosen based on a Voronoi construction. Variations on this method include picking grain orientations to produce a particular range of tilt angle (e.g., low angle grains), or picking grain centers to produce a log normal distribution of grain sizes [4, 12].

Many of the atomic simulations of sintering and grain growth dynamics – and many of the simulation studies in general on nanostructured materials – have focused on understanding how processes and rates at the nanometer scale differ from their counterparts in macroscopic-scale systems. At the macroscopic scale, six distinct mechanisms contribute to the sintering dynamics of crystalline particles: surface diffusion; lattice diffusion from the surface, from grain boundaries and through dislocations; vapor transport; and grain boundary diffusion. There is a much larger degree of surface curvature at the nanometer scale and a much higher ratio of interface to bulk atoms, both of which may lead to very different sintering mechanisms. The details regarding these differences, and some new and unexpected effects at the nanometer scale, have been revealed from computer simulations. Molecular dynamics simulations, for example, have been used to study surface energies, grain boundary mobility, and sintering of metal nanoparticle arrays at different temperatures [13, 14]. The results suggest that of the macroscopic-scale mechanisms associated with sintering, only surface and

grain boundary diffusion contribute significantly to nanometer-scale sintering dynamics, and that these two processes are accelerated due to the large interfacial forces. Simulations have also suggested three unconventional mechanisms that contribute to the early stages of nanometer-scale sintering: mechanical rotation, plastic deformation via dislocation generation and transmission, and amorphization of subcritical grains. This grain rotation mechanism has also been observed in structures with columnar grains, where simulations suggest that necessary changes in the grain shape during grain rotation in the nanostructure can be accommodated by diffusion either through the grain boundaries or through the grain interior [15].

In conventional metals with micron-scale grains, plastic deformation occurs by motion of dislocations. This dislocation motion can be inhibited by grain boundaries, which leads to the well-established Hall–Petch relation that the yield strength is proportional to the inverse square root of the grain size. However, there appears to be a threshold grain size below which materials become softer with decreasing grain size. This so-called inverse Hall–Petch behavior has been attributed to a transition from dislocation-mediated plastic deformation to grain boundary sliding for some critical grain size. This transition has been observed in several sets of molecular dynamics simulations that predict a transition grain size of about 10–15 nm in good agreement with experimental measurements. At the same time, the simulations have also revealed a rich and unanticipated set of dynamics near the threshold region that can be related to the fundamental properties of the bulk materials. These unique dynamics include an enhanced role of grain rotation (analogous to sintering dynamics), cooperative inter-grain motion, and formation of stacking faults via motion of partial dislocations across grains.

The deformation of strained nanocrystalline copper with grain sizes that average about 5 nm has been studied with molecular dynamics simulations. These simulations showed a material softening for small grain sizes, in agreement with experimental measurements. These simulations suggest that plastic deformation in the inverse Hall–Petch region occurs chiefly by grain boundary

sliding with dislocation motion having a minimal influence on deformation mechanisms. In related studies, molecular dynamics simulations have been used to understand the deformation of nanostructured nickel and copper with grain sizes ranging from 3.5 to 12 nm [5]. For grain sizes less than about 10 nm, deformation occurred mainly by grain boundary sliding, while for the larger grain sizes, deformation occurred by a combination of dislocation motion and grain boundary sliding. Detailed mechanisms of strain accommodation have been characterized; these include both single atom motion and correlated motion of several atoms, as well as stress-assisted free volume migration [16].

Molecular simulations of the deformation of columnar structures of aluminum have shown emission of partial dislocations during deformation that form at grain boundaries and triple junctions [17]. Atomic simulations have also suggested that these structures can be reabsorbed upon removal of the applied stress, which may contribute to an inability to observe dislocations experimentally in systems of this type after external stresses are released. In addition, near the grain size where plastic deformation transitions from dislocation-mediated plasticity to grain boundary sliding, the motion of single partial dislocations across nanograins during tensile loading has been observed in simulations. Without emission of a trailing partial dislocation, an intrinsic stacking fault is formed along the width of the nanograin. From the simulations, it appears that the formation of a low-energy ordered grain boundary drives emission of a full dislocation and the resulting absence of a stacking fault.

It has been argued that nucleation of the initial partial dislocation and the atomic rearrangement at the grain boundary associated with its emission sufficiently lowers the grain boundary energy such that emission of the trailing partial dislocation is not always needed to further relax the system [18]. Based on simulations of aluminum with a columnar nanostructure, it has been further suggested that the stacking fault width, and hence the intrinsic stacking fault energy, as determined by the distance between two partial dislocations, is the critical quantity that defines the transition from full to partial dislocation emission as grain

sizes approach the critical size for the onset of inverse Hall–Petch behavior [19]. On the other hand, it has been argued that the relation between the emission of partial dislocations does not correlate well with calculated stacking fault energies for nickel, copper, and aluminum. Instead, it has been suggested that full dynamics associated with the nucleation of a partial dislocation from a grain boundary must be considered, and therefore that the full planar fault energy, which includes the stable and unstable stacking fault energy as well as twin fault energies, must be used to understand and ultimately predict the relation between mechanical deformation and grain size.

Extensive simulations of crack propagation in nanostructured metals have also been carried out to better understand how fracture, fatigue, and toughness depend on grain scale [20–22]. These simulations have revealed crack propagation mechanisms that are similar to the plastic response of fully dense samples as well as key differences resulting from the presence of the crack tip. For example, in simulations of nanocrystalline nickel with grain sizes in the inverse Hall–Petch region, mode I crack propagation occurred by inter-grain decohesion via a mechanism involving coalescence of nanovoids that form in front of the crack tip. Plastic deformation leading to both full and partial dislocations was also observed in the neighboring grains.

Compared to their mechanical properties under tensile loading, much less is understood about the influence of grain size on the shock loading properties of nanostructured solids. Atomic simulations that have been carried out, however, suggest a strong coupling between nanostructure and shock loading dynamics, as well as unique and very important properties of shocked-loaded nanostructured metals. It has been noted, for example, that the mechanisms associated with the mechanical deformation of nanostructured metals depend strongly on pressure, temperature, and strain rate, and therefore, these materials may show ultrahigh strength under shock loading depending on the shock loading conditions and system [23]. The fast temperature and pressure rises associated with shock fronts freeze out deformation mechanisms that require diffusion. Similarly, production of

dislocations that requires nucleating events is inhibited. In the case of grain boundary accommodation, increasing the pressure results in an increase in the threshold stress for sliding plasticity. Similarly, the threshold for dislocation plasticity increases with increasing pressure because of an increase in the shear modulus with increasing pressure. Taking these effects into account, and assuming that the maximum in hardness as a function of grain size occurs where stress for sliding and dislocation plasticity are equal, it has been shown that ultrahigh hardness can be achieved by shock loading of nanostructured solids. These arguments have been validated by using molecular dynamics simulations to model the shock loading of nanostructured copper with different grain sizes. At relatively low shock velocities (i.e., low stresses), grain boundary sliding is observed, which results in a relatively low hardness value that increases with increasing grain size. At intermediate stresses, the hardness of the copper increases with increasing shock strength for all grain sizes, with a shift in the maximum hardness toward lower grain sizes compared to deformation at lower strain rates. This leads to a net increase in the maximum hardness of the material. At even higher stresses, simulations predict a drop in strength due to an increase in temperature and an associated increase in dislocation nucleation and motion.

It is clear from the discussion in the preceding sections that atomic simulations have provided new and exciting insights into the unique properties of nanosystems in general and nanostructured solids in particular. This remains a very active area of research within which molecular simulation will continue to provide new insights into relations between structural mechanical and thermodynamic properties.

S

Cross-References

- ▶ Computational Study of Nanomaterials: From Large-scale Atomistic Simulations to Mesoscopic Modeling
- ▶ Molecular Dynamics Simulations of Nano-biomaterials
- ▶ Nanomechanical Properties of Nanostructures

References

1. Brenner, D.W., Shenderova, O.A., Areshkin, D.A.: Quantum-based analytic interatomic forces and materials simulation. In: Lipkowitz, K.B., Boyd, D.B. (eds.) *Reviews in Computational Chemistry*, pp. 213–245. VCH, New York (1998)
2. Brenner, D.W., Shenderova, O.A., Schall, J.D., Areshkin, D.A., Adiga, S., Harrison, J.A., Stuart, S.J.: Contributions of molecular modeling to nanometer-scale science and technology, Chapter 24. In: Goddard, W., Brenner, D., Lyshevski, S., Iafrate, G. (eds.) *Nanoscience, Engineering and Technology Handbook*. CRC Press, Boca Raton (2002)
3. Wolf, D., Yamakov, V., Phillpot, S.R., Mukherjee, A., Gleiter, H.: Deformation of nanocrystalline materials by molecular dynamics simulations: relation to experiment? *Acta Mater.* **53**, 1 (2005)
4. Kumar, K.S., Van Swygenhoven, H., Suresh, S.: Mechanical behavior of nanocrystalline metals and alloys. *Acta Mater.* **51**, 5743 (2003)
5. Van Swygenhoven, H., Spaczek, M., Caro, A., Farkas, D.: Competing plastic deformation mechanisms in nanophase metals. *Phys. Rev. B* **60**, 22 (1999)
6. Yanakov, V., Wolf, D., Salazar, M., Phillpot, S.R., Gleiter, H.: Length-scale effects in the nucleation of extended dislocations in nanocrystalline Al by molecular-dynamics simulation. *Acta Mater.* **49**, 2713 (2001)
7. Phillpot, S.R., Wolf, D., Gleiter, H.: Molecular-dynamics study of the synthesis and characterization of a fully dense, three-dimensional nanocrystalline material. *J. Appl. Phys.* **78**, 847 (1995)
8. Phillpot, S.R., Wolf, D., Gleiter, H.: A structural model for grain boundaries in nanocrystalline materials. *Scripta Metall. Mater.* **33**, 1245 (1995)
9. Zhang, Y.W., Liu, P., Lu, C.: Molecular dynamics simulations of the preparation and deformation of nanocrystalline copper. *Acta Mater.* **52**, 5105 (2004)
10. Kodiyalam, S., Kalia, R., Nakano, A., Vashashta, P.: Multiple grains in nanocrystals: effect of initial shape and size on transformed structures under pressure. *Phys. Rev. Lett.* **93**, 203401 (2004)
11. Chatterjee, A., Kalia, R.K., Nakano, A., Omelchenko, A., Tsuruta, K., Vashishta, P., Loong, C.-K., Winterer, M., Klein, S.: Sintering, structure, and mechanical properties of nanophase SiC: a molecular-dynamics and neutron scattering study. *Appl. Phys. Lett.* **77**, 1132 (2000)
12. Schiottz, J., Di Tolla, F.D., Jacobsen, K.W.: Softening of nanocrystalline metals at very small grain sizes. *Nature* **391**, 1223 (1998)
13. Zeng, P., Zajac, S., Clapp, P.C., Rifkin, J.A.: Nanoparticle sintering simulations. *Mater. Sci. Eng. A* **252**, 301 (1998)
14. Xiao, S., Hu, W.: Molecular dynamics simulations of grain growth in nanocrystalline Ag. *J. Cryst. Growth* **286**, 512 (2006)
15. Haslam, A.F., Phillpot, S.R., Wolf, D., Moldovan, D., Gleiter, H.: Mechanisms of grain growth in nanocrystalline fcc metals by molecular-dynamics simulation. *Mater. Sci. Eng. A* **318**, 293 (2001)
16. Van Swygenhoven, H., Derlet, P.M.: Grain boundary sliding in nanocrystalline fcc metals. *Phys. Rev. B* **64**, 224105 (2001)
17. Yamakov, V., Wolf, D., Salazar, M., Phillpot, S.R., Gleiter, H.: Length scale effects in the nucleation of extended dislocations in nanocrystalline Al by molecular dynamics simulation. *Acta Mater.* **49**, 2713 (2001)
18. Van Swygenhoven, H., Derlet, P.M., Froseth, A.G.: Stacking fault energies and slip in nanocrystalline metals. *Nat. Mater.* **3**, 399 (2004)
19. Yamakov, V., Wolf, D., Phillpot, S.R., Mukherjee, A.K., Gleiter, H.: Deformation mechanism map for nanocrystalline metals by molecular dynamics simulation. *Nat. Mater.* **3**, 43 (2004)
20. Latapie, A., Farkas, D.: Molecular dynamics investigation of the fracture behavior of nanocrystalline α -Fe. *Phys. Rev. B* **69**, 134110 (2004)
21. Farkas, D., Van Petegem, S., Derlet, P.M., Van Swygenhoven, H.: Dislocation activity in nano-void formation near crack tips in nanocrystalline Ni. *Acta Mater.* **53**, 3115 (2005)
22. Farkas, D., Sillemann, M., Hyde, B.: Atomistic mechanisms of fatigue in nanocrystalline metals. *Phys. Rev. Lett.* **94**, 165502 (2005)
23. Bringa, E.M., Caro, A., Wang, Y., Victoria, M., McNamee, J.M., Remington, B.A., Smith, R.F., Torralva, B.R., Van Swygenhoven, H.: Ultrahigh strength in nanocrystalline materials under shock loading. *Science* **309**, 1838 (2005)

Simulations of Oxide/Polymer Hybrids

Maria Ilenia Saba and Alessandro Mattoni
 Istituto Officina dei Materiali, Consiglio
 Nazionale delle Ricerche, CNR-IOM Cagliari
 (SLACS), Monserrato, CA, Italy

Synonyms

Atomistic simulations; Molecular dynamics;
 Organic/inorganic interactions

Definitions

Polymer/oxide hybrids are systems composed mainly by polymers and oxides.

Oxides are intended here as a class of inorganic solid compounds having at least one oxygen and one other element (e.g., metal or semimetal) in their chemical formula. They typically contain the oxygen anion in the -2 oxidation state. Oxides are abundant in nature; most of the Earth's crust consists of solid oxides, due to the oxidation of the elements in air or in water.

Polymers are a class of natural or synthetic chemical compounds formed by the repetition of small units (monomers). Polymers formed by organic units are named organic polymers. An important class of polymers are the conjugated ones, having a backbone of alternating double and single bonds that provides conductive properties. Polymer/oxide systems have applications in bioengineering, textiles, magnetism, optoelectronics, and photovoltaics.

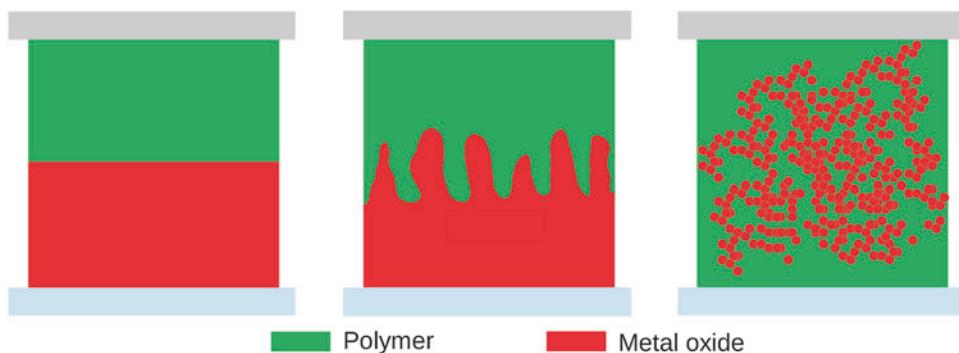
Molecular dynamics is a computational technique that consists in calculating the classical trajectories of a set of interacting atoms representing the material of interest. The trajectories obey the Newton's equation of motions ($\vec{F} = m \vec{a}$), where the interatomic forces \vec{F} are calculated as the derivative of a suitable potential. In model potential molecular dynamics, the potential does not depend explicitly on the electronic degrees of freedom, but it is an empirical function of the nuclei positions with parameters that are calibrated to reproduce a set of physical properties of the material. Model potential molecular dynamics is a valuable tool to study hybrid polymer/oxide systems since it makes affordable the study of large-scale atomistic models with complex morphologies while retaining an atomistic (i.e., fundamental) description of the material.

Technological Relevance of Polymer/Oxide Hybrids

Polymer/oxide materials have high technological impact in textiles, bioengineering, molecular electronics, catalysis, magnetism, optoelectronics, and photovoltaics [1]. Their hybrid nature makes possible to join, at a low production cost, the tailororable properties and the flexibility of the

organic polymers with the thermal and mechanical stability or the good transport properties of the inorganic components. A large number of synthetic or natural polymers can be used in hybrids, including organic, inorganic, and biopolymers. In technical textiles, for example, organic polymers such as polyolefins, nylons, polyesters, and polyurethanes are widespread used because of easy fabrication, processability, durability, and relatively low cost [2]. The inorganic particles improve the tensile strength or stiffness via reinforcement mechanisms described by theories for nanocomposites. The addition of inorganic particles (such as metal oxides and silica) not only provides mechanical and thermal stability but also new functionalities that depend on the chemical nature, the structure, the size, and the crystallinity of the inorganic nanoparticles. For example, titania (TiO_2) can be mixed with poly(amide-imide) forming composite membranes for gas separation or with poly(3-hexylthiophene) for photovoltaics; vanadium oxide (V_2O_5) can be used with poly(3,4-ethylenedioxythiophene) for cathode materials for rechargeable lithium batteries; zinc oxide (ZnO) particles with polystyrene resin are useful for electrical applications and with poly(3-hexylthiophene) for optoelectronics and photovoltaics; alumina (Al_2O_3) mixed with polycarbonate can be applied as optical materials; and Fe_3O_4 with poly(vinylidene difluoride) is useful for magnetic nanocomposite and $\gamma-Fe_2O_3$ with polyimide for superparamagnetic films. Finally, silica (SiO_2) can be combined with polycaprolactone, polyimide, poly(methyl methacrylate), polyethyl acrylate, poly(p-phenylene vinylene), and polycarbonate giving rise to hybrids applicable for bone-bioerodible and skeletal tissue repair, microelectronics, dental application and optical devices, catalysis support and chromatography, nonlinear optical material for optical waveguides, and abrasion-resistant coating, respectively [1–3].

Among polymers, conductive ones are the most relevant when considering hybrids for optoelectronics and photovoltaics. They are characterized by a backbone formed by conjugated chemical groups with alternating double and single bonds, along which the charge carriers



Simulations of Oxide/Polymer Hybrids, Fig. 1 Microstructure of polymer/metal oxide hybrids: bilayers (*left*), polymer/porous metal oxide bulk heterojunction (*center*), and polymer/nanocrystal bulk heterojunction (*right*)

(in most cases holes) can diffuse easily. One of the most commonly used is poly(3-hexylthiophene) (P3HT) that has good hole mobility ($0.1 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$) and thermal stability. The backbone of P3HT consists of π -conjugated thiophenes, responsible for the good intrachain hole mobility, and alkyl side chains, providing polymer solubility and processability in several solvents.

Concerning oxides, the most used for hybrid polymer/oxide composites are ZnO, TiO₂, and SiO₂. Zinc oxide (ZnO) is a wide band gap semiconductor (3.37 eV) that crystallizes in two main forms: hexagonal wurtzite (the most common at ambient conditions) and cubic zincblende. ZnO has very good electron mobility ($205 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$); it is nontoxic, and it can be grown in a variety of highly crystalline nanostructures which are commonly used as electron acceptors. Nanorods with (10 $\overline{1}$ 0) nonpolar surfaces are commonly used in hybrid bulk heterojunctions for photovoltaics [4].

Titanium dioxide (TiO₂) is a semiconductor used as sensor, pigment, and photocatalyst and in solar cell for the production of hydrogen and energy. TiO₂ crystallizes in three major structures: tetragonal rutile (the most common and stable), anatase, and rhombohedral brookite. The most stable crystalline surfaces are the (110) for the rutile phase and the (101) for the anatase one.

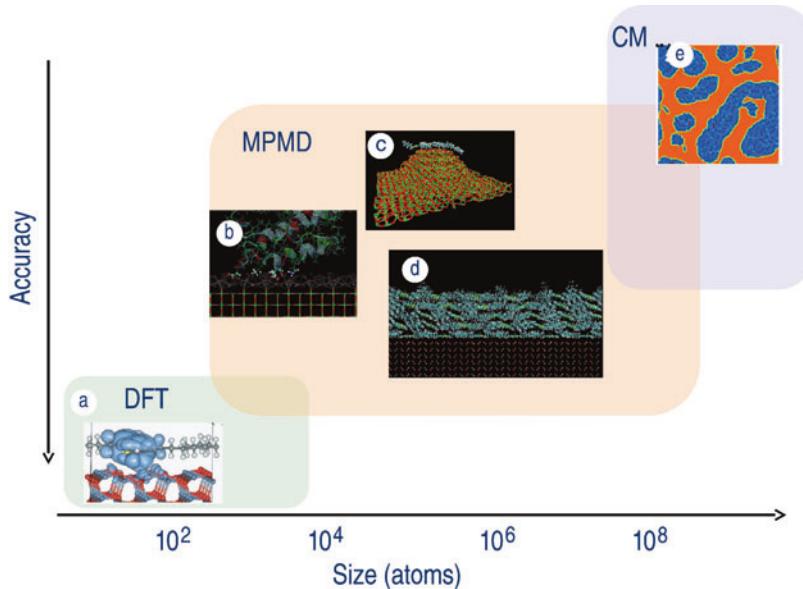
Silica (SiO₂) is one of the most abundant oxides in nature. In addition to its amorphous disordered phases, it has a number of crystalline structures (polymorphs). The only stable

crystalline structure under normal thermodynamic conditions is the α -quartz that is formed by SiO₄ tetrahedra linked by shared corner oxygen atoms. Due to its stability and biocompatibility, SiO₂ is largely used in biomedical applications.

Atomistic Modeling of Hybrid Systems

The microstructure of polymer/metal oxide hybrids can be very different depending on the synthesis methods, on the technological applications, and on the compatibility between the organic and the inorganic components. Simple planar interfaces (also called bilayers; see Fig. 1 left) are sometimes used for sensors, optoelectronic devices, or coatings. More complex microstructure can be found in hybrids for structural or optoelectronic applications. For example, in the bulk heterojunction architectures used in photovoltaics, the polymer is infiltrated within a mesoporous oxide scaffold (Fig. 1 center) or mixed with oxide nanoparticles (right) so forming two interpenetrated networks of polymer and oxide. Such kind of microstructures provides a large interface-to-volume ratio necessary to optimize the light absorption and the collection of carriers through percolation paths to the contacts.

Due to the complex polymer/metal oxide microstructure, the overall macroscopic properties of the hybrid depend on several physical and chemical processes occurring at different length



Simulations of Oxide/Polymer Hybrids,

Fig. 2 Accuracy versus size in atomistic simulation of polymer/metal oxide hybrids: (a) ab initio P3HT/ZnO system (Adapted with permission from Dag et al., *Nano Letters* 2008 8 (12), 4185–4190. Copyright (2008) American Chemical Society), (b) protein adsorbed on TiO₂ surface studied by molecular dynamics (Adapted with permission from Kang et al. *J. Phys. Chem. C* 2010 114 (34), 14496–14502. Copyright (2010) American Chemical Society), (c) snapshot of MPMD simulation of a P3HT polymer chain on TiO₂ surface (Adapted with permission

from Melis et al. *J. Phys. Chem. C* 2010 114 (8), 3401–3406, Copyright (2010) American Chemical Society), (d) P3HT/ZnO interface studied by MPMD (Adapted with permission from Saba et al. *J. Phys. Chem. C* 2014 118 (9), 4687–4694. Copyright (2014) American Chemical Society), (e) snapshot of the time evolution of Janus nanorod-filled binary polymer mixture investigated by using a cell dynamics system (CDS) method (Adapted with permission from Li et al. *Macromolecules*, 2013, 46 (18), pp 7465–7476. Copyright (2013) American Chemical Society)

and timescales: at the mesoscopic scale, where the material can be described as a continuum blend, and down to the atomic scale, where chemical bonds, electrostatic, and dispersive forces between atomic constituents control, for example, the interface properties.

Atomistic methods are necessary to model predictively the physical properties of hybrid materials and are complementary to the experimental characterization, as well as the continuum modeling approaches. Among the other atomistic methods, the model potential molecular dynamics (MPMD) is a powerful tool to study the complex microstructure of hybrid materials since, while retaining an atomistic (i.e., fundamental) description, it makes affordable the study of relatively large systems (typically composed by 10⁴–10⁶ atoms). At variance with ab initio molecular dynamics, that is typically limited to systems as

small as 10²–10³ atoms because of the high computational cost associated to the explicit treatment of electrons, in MPMD the electronic degrees of freedom are integrated into an effective interatomic potential that only depends on the nuclei positions. The accuracy and the transferability of the method must be validated in each case, but the computational workload is highly reduced. It typically scales linearly with the number of atoms (order-N method), making possible to study the thermodynamics and the microstructure evolution of systems as large as 10¹–10³ nm (up to several millions of atoms) over the 10¹–10³ ns timescale (see Fig. 2). MPMD is a suitable method to study hybrid systems where a large number of atoms are necessary to represent the structural disorder induced by macromolecules and interfaces. Furthermore, long-range electrostatic and dispersive interactions that are

important in hybrid materials are easily incorporated into MPMD by using two-body Coulomb and Lennard-Jones-type potentials, respectively. The drawback of the method is that the accurate description of interatomic forces is not guaranteed and careful calibrations against experimental data or ab initio calculations are necessary.

The Force Fields for Oxide/Polymer Systems

The interatomic potential for hybrid polymer (P) / oxide (X) systems to be used in model potential molecular dynamics can be written in most cases as the sum of organic–organic, organic–inorganic, and inorganic–inorganic interactions:

$$U = U_{PP} + U_{PX} + U_{XX}$$

The basic requirement of any interatomic potential for hybrid systems is to provide a good description of the separate organic and inorganic phases.

Several potentials exist for the U_{PP} term that describes the chemical bonding between atoms of the organic polymers, as well as intermolecular interactions. A possible classification of the U_{PP} force fields is based on the number of atomic degrees of freedom simulated in the calculation. All-atom force fields treat explicitly all the atoms of the system, including hydrogen; this is the case of AMBER (Assisted Model Building with Energy Refinement) [5], CHARMM (Chemistry at HARvard Molecular Mechanics) [6], and OPLS (Optimized Potentials for Liquid Simulations) [7] force fields. United-atom force fields (such as Gromos (GROningen MOlecular Simulation package) [8]) do not include explicit representation of nonpolar hydrogen atoms; only polar hydrogens are included in the force field definition. Finally, in the coarse-grained force fields, specific chemical groups are replaced by a single or a few pseudo-atoms representing the collective behavior of the group. In this way, the number of degrees of freedom of the simulation is smaller and the computational workload is strongly reduced making affordable larger systems and timescale. An example of coarse-grained force field developed for biomolecular systems is the Martini force field [9]. It has been parameterized

to reproduce the partitioning free energies between polar and apolar phases of a large number of chemical compounds.

The AMBER force field is the typical all-atom classical force field to study organic materials. It takes into account bonded and nonbonded interactions [5]. The former are necessary to model the chemical bonds between atoms. The list of interacting atoms must be specified according to the topology of the molecules and it is fixed during the simulation (molecular mechanics). Bonding interactions consist of three contributions [5]:

$$U_{\text{bonded}} = U_{\text{bonds}} + U_{\text{angles}} + U_{\text{dihedrals}} \quad (1)$$

$U_{\text{bonds}} = \sum_{ij} \frac{1}{2} K_{ij}^b (r_{ij} - r_{ij}^0)^2$ describes the two-body elastic part of the bonding. K_{ij}^b is the constant of the force, r_{ij}^0 is the equilibrium distance between two atoms ij , and r_{ij} is the corresponding instantaneous distance of the bonded atoms (Fig. 3 top left).

$U_{\text{angle}} = \sum_{ijk} \frac{1}{2} K_{ijk}^a (\theta_{ijk} - \theta_{ijk}^0)^2$ describes the energy cost of the deformation of the angle formed by three adjacent atoms ijk . K_{ijk}^a and θ_{ijk}^0 are the stiffness and the equilibrium angle, and θ_{ijk} is the instantaneous bond that they form (Fig. 3 top right).

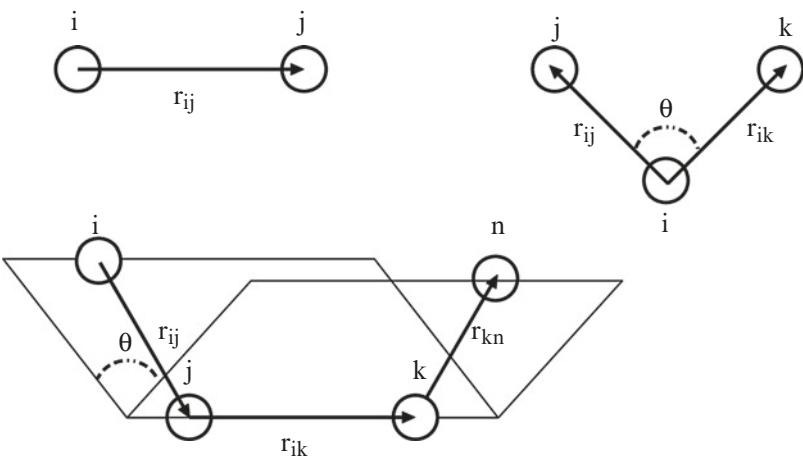
$U_{\text{dihedral}} = \sum_{ijkn} \frac{1}{2} V_{ijkn}^\phi (1 + \cos(n_{ijkn} \phi_{ijkn} - \phi_{ijkn}^0))$ describes the interaction arising from torsional forces in molecules. They are sometimes referred to as torsion potentials and require the specification of four atomic positions representing the instantaneous angle between the plane defined by the atoms ijk and that defined by the atoms jkn . V_{ijkn}^ϕ , ϕ_{ijkn}^0 , and n_{ijkn} are the stiffness, the equilibrium dihedral angle, and the multiplicity, respectively, associated with the torsion involving $ijkn$ types of atoms (Fig. 3 bottom).

Improper dihedrals, involving nonconsecutive atoms, can also be considered and are generally used to enforce planarity around sp^2 central atoms [5].

The bonded energy U_{bonded} prevents bond breaking, and it is suitable for describing phenomena in which the chemical groups do not dissociate during the simulation.

**Simulations of Oxide/
Polymer Hybrids,**

Fig. 3 Bonding (top left), angular (top right), and dihedral (bottom) interaction between two, three, and four atoms



The nonbonded terms make it possible to establish interactions between atoms that are not chemically bonded or that are separated by three or more bonds. It is modeled as a sum between the Coulombic interactions between atomic partial charges q_i and the van der Waals forces described by the Lennard-Jones 12–6 potential:

$$U_{\text{non-bonded}} = U_{vdW} + U_{\text{Coul}} = \sum 4\epsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}} - \frac{\sigma_{ij}^6}{r_{ij}^6} \right) + \sum \frac{q_i q_j}{4\pi\epsilon r_{ij}} \quad (2)$$

where ϵ is the depth of the van der Waals potential well, σ is the finite distance at which the interparticle potential is zero, and r_{ij} is the distance between the particles and ϵ is the dielectric constant. The r_{ij}^{-12} term describes the Pauli repulsion at short distances due to electron overlapping orbitals, while the attractive long-range r_{ij}^{-6} term corresponds to the interactions between fluctuating dipoles (dispersion forces). Other U_{PP} potentials such as CHARMM and OPLS force fields have a functional form very similar to that of AMBER.

There exists a second generation of interatomic potentials with higher-order and nondiagonal mixing terms for the bonded interactions [10]. Among them, there are CFF [11], UFF [12], and the MM series force fields of Allinger [13]. Derived from CFF, COMPASS [14] force

field combines ab initio results and the empirical fitting of condensed-phase properties to obtain parameters for valence, charge, and van der Waals terms.

Another classification for the force fields is related to the way in which the atomic charge is treated. In first-generation force fields, the charge is kept fixed, usually centered on atoms, without taking into account the electrostatic environment and the electronic polarization. In the last years, a new generation of force field incorporating models for polarizability has been proposed. The most studied approach is the point dipole method [10], included, for example, in the AMOEBA (Atomic Multipole Optimized Energies for Biomolecular Applications) [15] force field, where polarization energy is achieved by the interactions between static charges and their induced dipole moments [10]. In the fluctuating charge model (e.g., included in the UFF and CHARMM force fields), a charge flows between atoms equalizing their electronegativities. In this way, a molecule's charge-state distribution is coupled to its environment and the effects of polarization are included [10]. In Drude oscillator methods (or shell methods), the electronic polarizability is included adding an extra charged site to each ion connected via a spring. The magnitude of both charges is fixed, and the electronic polarization is mimicked by relative displacement of both charges due to an external electrostatic field [10].

Concerning the inorganic–inorganic term U_{XX} , it is usually of nonbonded type. For example, for ZnO and TiO₂, U_{XX} can be modeled by adding to the long-range electrostatic interactions, a two-body Buckingham potential [16, 17]

$$U_{\text{buck}} = \sum_{ij} A \exp\left(-\frac{r_{ij}}{B}\right) - \frac{C}{r_{ij}^6} \quad (3)$$

where A , B , and C are fitting parameters. The atomic partial charges used in this approach are fixed and polarizability effects can be included, for example, by using shell methods.

Many-body effects can be included by environment-dependent potentials, such as the reactive force field (ReaxFF) [18]. It is a bond-order interaction model with two-body, three-body, and four-body short-range interaction terms. ReaxFF have been applied to ZnO [18] and TiO₂ [19]. It allows the redistribution of charges, can simulate the breaking and reforming of bonds, and can reproduce the structures and mechanical properties of condensed phases but requires a high computational cost and a very large number of fitting parameters.

As for the modeling of SiO₂, several force field have been applied, in order to take into account its significant ionic contribution and the strong, directional covalent bond. Woodcock et al. were the first to perform simulations on vitreous silica using an ionic pair potential [20], while Feuston and Garofalini introduced the representation of the covalent bonding by a three-body interatomic potentials [21]. Vessal et al. used full ionic charges and developed two different models to take into account the directionality of the Si–O bond [22]. In both cases, they used a two-body (a Buckingham potential) plus a three-body interaction (called screened and truncated) [22]. The Beest, Kramer, and van Santen (BKS) model uses partial charges to simulate partial covalent bonds and short-range interactions of the Buckingham form [23]. Morse potentials [24, 25] as well as ReaxFF [26] have been developed for silica.

The last potential contribution of hybrid polymer/oxides is the U_{PX} that contains the cross atomic interactions between polymer and the oxide. In most cases, there are not covalent

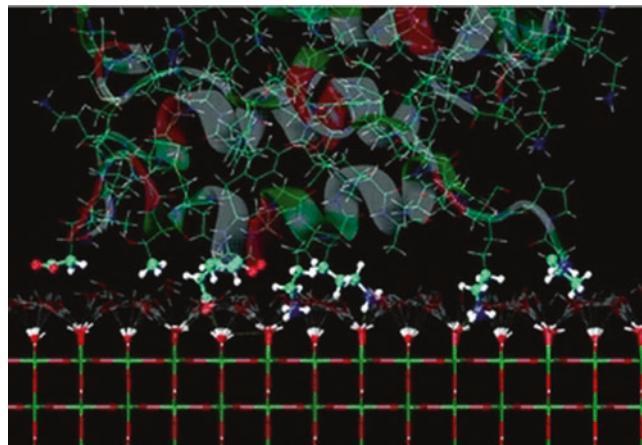
bonds between the polymer and the oxide (unless specific ligands with anchoring groups are present). Under these conditions, the interaction can be treated by nonbonding forces of dispersive or electrostatic nature, provided that a proper fit of the charges is used. U_{PX} models of nonbonded type can be successfully applied to reproduce the interaction of aromatic rings (e.g., thiophene and benzenes within conjugated polymers) with an oxide surface (e.g., ZnO, TiO₂, SiO₂). Also in the presence of chemical ionic binding between the polymer and the oxide, the same approach can be effective. Most organic ligands such as fatty acids (e.g., oleic acid) have a carboxylic COOH terminal group able to anchor to the oxide surface. MPMD simulations are able to reproduce monodentate and bidentate COO–X anchoring modes of protonated and deprotonated molecules for several ligands. The same strategy can be successfully employed for POOH–X interactions and heterocycles containing oxygen (e.g., THF [27]) or nitrogen (such as pyridines on titania [28], for which an effective N–Ti binding is reproduced), as well as for organometallic compounds such as zinc phthalocyanine on ZnO [29].

As for the computational cost of the above force fields, the nonbonded electrostatic interactions are in most cases the most expensive, with the direct computation time proportional to the square of the number of atoms N^2 . An efficient method to spare computational time is represented by the Ewald summation. In this method from each point charge spreads out radially, up to a cutoff distance, a Gaussian charge distribution of the same magnitude and opposite sign. The interaction is separated into short-range interaction, representing the screening interaction between neighboring charges and calculated in the real space, and long-range one, representing the canceling charge distribution of the same sign as the original charge calculated in the reciprocal space [30]. In this way, the time calculation is proportional to $N^{3/2}$. By calculating the Ewald summation using the fast Fourier transform (FFT) technique (particle-mesh Ewald method), the computational workload can scale as $N \log(N)$.

Concerning the cost of the U_{PP} part, in most cases, its terms (even in the case of environment-

Simulations of Oxide/ Polymer Hybrids,

Fig. 4 Protein adsorption on a hydroxylated TiO_2 surface (Adapted with permission from Kang et al. J. Phys. Chem. C 2010 114 (34), 14496–14502. Copyright (2010) American Chemical Society)



dependent contributions) decrease fast with interatomic distances and they are set to zero beyond a cutoff length. Accordingly, the overall workload of the force fields $U = U_{PP} + U_{PX} + U_{XX}$ scales linearly with the number of atoms N in the hybrid system.

Applications of Atomistic Methods to Hybrid Systems

Atomistic simulations can be applied to study a large number of physical properties of interest for hybrids such as adhesion and interface energies, molecular order and crystallinity, diffusivity and assembling phenomena, vibrational properties, solvent effects, and so on. In most atomistic studies, the central problem is related to the modeling of the polymer/oxide interface from planar interfaces to bulk heterojunctions. Below we illustrate a few paradigmatic examples for biology, photovoltaics, and optoelectronics.

Atomistic Studies for Biological Applications

An important problem related to hybrid systems for biology is the interaction of biopolymers (such as proteins) with oxides and, in particular, with TiO_2 . Titanium has been extensively employed for medical implants such as dental ones, artificial joints, and blood-contacting devices. It is regarded as a bioactive material with good biocompatibility, which mainly derives from the presence of the surface oxide film that protects

the metal from further oxidation [31]. Biocompatibility requires to selectively control the hybrid interactions and the adsorption between the biopolymers and the implant surfaces. The interactions among the proteins, the TiO_2 surface, and the solvent, as well as the dynamic mechanism of the adsorption/desorption process, are the key points to design implants with optimal bioactivity and antifouling properties (i.e., inhibiting the accumulation of undesired materials on wetted surfaces through nonspecific interactions).

Many proteins such as albumin, fibronectin, and laminin have been found to be able to adsorb onto titanium dioxide surfaces. For example, Kang et al. [31] have studied by model potential molecular dynamics the role of hydroxylation in the absorption of human serum albumin onto TiO_2 in water environment (Fig. 4). The model potential adopted was the AMBER force field for the protein, the Buckingham potentials for the titania, and the Lennard-Jones and Coulombic interactions for the hybrid interactions. By molecular dynamics simulations, it was shown that the hydroxylated surface has a stronger interaction with the protein and provides a larger number of adsorbed residues with respect to the non-hydroxylated case. The hydroxyl modification makes it easier for the protein to move closer to surface by reducing the hydrogen bonds between the surface and the water molecules.

The same model potential for TiO_2 has been used in the study of the titania interaction with the biopolymers surrounding the cell membranes.

In particular, Ham et al. by using the OPLS force field investigated the adhesion on titania of a glycopeptoid that mimics the composition and abilities of glycocalyx [32]. This is a glycoprotein–polysaccharide with antifouling properties that surrounds the cell membranes of some bacteria, epithelia, and other cells. The computational results revealed an aqueous interface enriched in highly hydrated saccharide residues with the formation of a high number of hydrogen bonds with water molecules possibly contributing to fouling resistance.

Biopolymeric films on metal oxides have been also exploited in preventing blood coagulation and thrombosis. A class of copolymers based on poly(L-lysine)-g-poly(ethylene glycol) (PLL-g-PEG) was found to spontaneously adsorb from aqueous solutions onto several metal oxide surfaces, such as TiO_2 , $\text{Si}_{0.4}\text{Ti}_{0.6}\text{O}_2$, and Nb_2O_5 [33]. The resulting adsorbed layers are highly effective in reducing the adsorption both of blood serum and of individual proteins such as fibrinogen, which is known to play a major role in the cascade of events that lead to biomaterial-surface-induced blood coagulation.

Though classical molecular dynamics is the method of choice for most atomistic investigations in biophysics, the applications to polymer/oxide systems are relatively few. A rapid growth of the field is expected in the next years.

Atomistic Studies for Photovoltaic and Optoelectronic

A class of polymer/oxide hybrid films formed by organic conductive polymers (such as P3HT) and metal oxides (typically TiO_2 or ZnO) have attracted a large interest as candidates for environmentally friendly and solution-processable films for low-cost photovoltaics. In these hybrids, the polymer and the metal oxide form a donor–acceptor pair with photovoltaic properties in which the polymer absorbs the sunlight, donates the electrons to the oxide, and transports the holes to the contact. The electrical and optical properties of the system are very susceptible to the microstructure and to the crystalline assembling of the molecular constituents. There is in fact a strong dependence of the intermolecular

electronic properties (e.g., transfer integral) on the ordered stacking of the polymer chains. When the polymer order is lost, as is the case of amorphous polymers, an increase of resistivity and a blue shift of the light absorption are observed. Accordingly, the polymer order and crystallinity at the polymer/oxide interface has been the subject of intense investigation by atomistic methods.

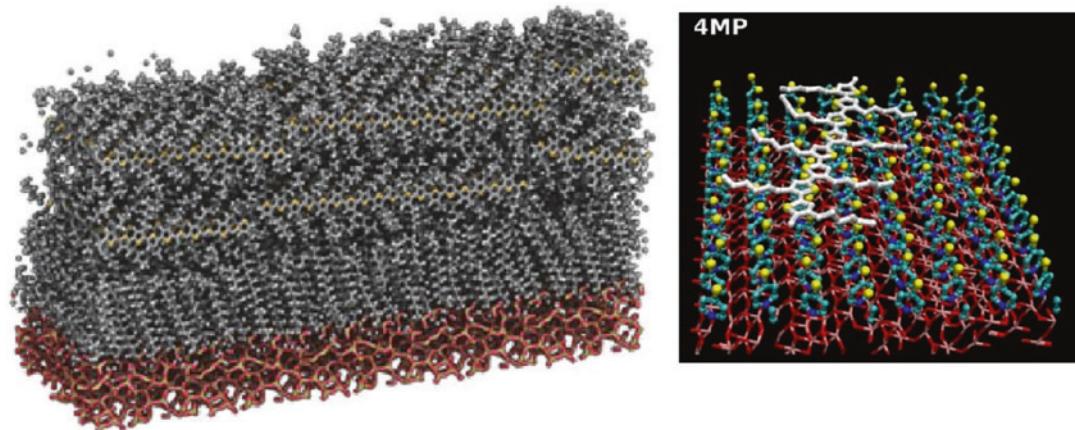
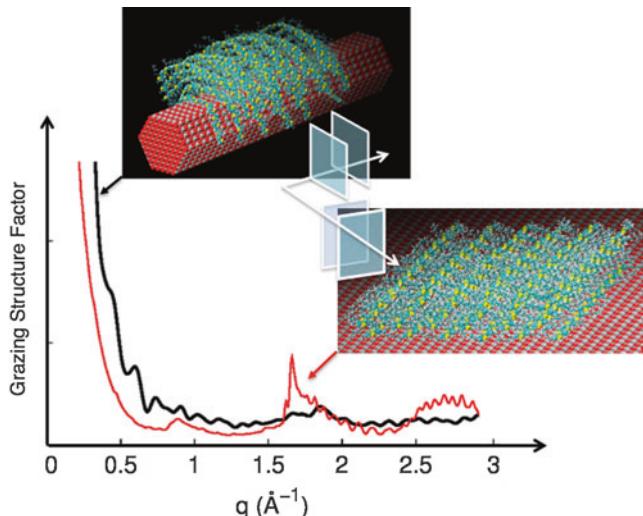
In this context, one example is that of P3HT/ ZnO hybrid interface that has been studied to understand the spontaneous formation of disordered polymer close to the interface with the oxide [34]. By atomistic simulations (based on AMBER, Buckingham, and Coulombic plus dispersive model potentials), it has been shown that the polymer/ ZnO adhesion is driven by the electrostatic interactions between the polymer backbones and the oxide. On planar ZnO substrates (Fig. 5 right inset), though some disorder is present, the polymer films can be crystalline with predominance of the face-on molecular orientation. However, the local curvature of the oxide substrate can induce the bending of the polymer backbones [35], perturbing the molecular film with the possible formation of amorphous layers (Fig. 5 top inset). Specific methods to analyze the degree of order/disorder in the atomistic models are necessary, such as the structure factor analysis. It has been applied to the P3HT/ ZnO surface showing a decrease of the diffraction peak of the polymer backbones with the substrate curvature (Fig. 5), in agreement with experimental evidences.

A further important problem of interest for atomistic simulation is related to the use of interfacial layers between the oxide and the polymers. The use of the interlayer represents a versatile approach to engineer and design hybrid interfaces with controlled order and improved properties. For example, the case of P3HT polymer on SiO_2 that is relevant in the field of hybrid and organic transistors has been studied by Meredig et al. (see left panel of Fig. 6). By using the COMPASS force field, they have shown that the presence of a disorder self-assembled monolayer (SAM) of octadecyltrichlorosilane (OTS) between the polymer and the silica substrate [36] induces a more stable edge-on orientation for P3HT.

**Simulations of Oxide/
Polymer Hybrids,**

Fig. 5 Planar and curved P3HT/ZnO interfaces and comparison between their structure factors (Adapted with permission from Saba et al. J. Phys. Chem. C 2014 118 (9), 4687–4694.

Copyright (2014) American Chemical Society)



Simulations of Oxide/Polymer Hybrids, Fig. 6 *Left:* edge-on P3HT chains on amorphous SiO₂ surface with a SAM interlayer of OTS molecules in between (Adapted with permission from Meredig et al. ACS Nano 2009 3 (10), 2881–2886. Copyright (2009) American Chemical

Society) *Right:* example of ternary P3HT/4MP/TiO₂ system (Adapted with permission from Malloci et al. J. Phys. Chem. C 2013, 117 (27), 13894–13901. Copyright (2013) American Chemical Society)

S

Interfacial layers are also useful to improve the mobility of polymer molecules on the oxide surface. Such a mobility enhancement is expected to improve the quality and the amount of polymer/oxide interface area. For example, atomistic simulations at the P3HT/TiO₂ interface showed a highly increased polymer mobility in the presence of self-assembled interlayers of 4-mercaptopuridines (4MP) [28] (see Fig. 6 right). These molecules (similar to pyridines that are typically used as exchange ligand for titania

nano-particles) are able to efficiently bind to the titania forming a crystalline monolayer that exposes a flat surface of thiol groups to the polymer. This improves the ability of the polymer to infiltrate within mesoporous titania.

Interlayers are also important to enhance optical absorption, charge separation, and injection at the polymer/oxide interface. For these purposes, the molecular interlayer can be formed by optically active molecules such as dyes. There exist several examples in the field of solid-state

dye-sensitized solar cells in which a polymer (P3HT) is mixed with an oxide (titania or zinc oxide) sensitized by organic absorbing molecules. An example of atomistic simulation in this context studies the role of phthalocyanines in P3HT/ZnO systems [37]. A multi-scale combination of atomistic methods, including model potential molecular dynamics, has been applied to understand the thermodynamic stability of the system and its optoelectronics properties. The molecules are found to bind in parallel configuration to the substrate through Zn-O interactions, and they self-assemble into oriented stripes that cover the oxide surface. The formation of the ZnPc film has been shown to be beneficial for absorption, charge injection, and energy level alignment [37].

Model potential molecular dynamics methods are emerging as powerful tools for hybrid polymer/oxide systems. It is expected that its role in the design of novel hybrid materials will grow together with the increase of computer powers and the development of more advanced interatomic potentials. The combination of model potentials to first-principle methods, on the one hand, and to coarse-grained or continuum methods, on the other, will further extend the relevance of the atomistic simulations of polymer/oxide systems.

Cross-References

- [Ab Initio DFT Simulations of Nanostructures](#)
- [Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling](#)
- [Hybrid Solar Cells](#)
- [Molecular Dynamics Simulations of Nano-biomaterials](#)
- [Nanomaterials for Excitonic Solar Cells](#)

References

1. Sanchez, C., Beatriz, J., Belleville, P., Popall, M.: Applications of hybrid organic-inorganic nanocomposites. *J. Mater. Chem.* **15**, 3559–3592 (2005)
2. Jeon, I.-Y., Baek, J.-B.: Nanocomposites derived from polymers and inorganic nanoparticles. *Materials* **3**, 3654–3674 (2010)
3. Camargo, P.H.C., Satyanarayana, K.G., Wypych, F.: Nanocomposites: synthesis, structure, properties and new application opportunities. *Mater. Res.* **12**(1), 1–39 (2009)
4. Bouclé, J., Ackermann, J.: Solid-state dye-sensitized and bulk hetero-junction solar cells using TiO₂ and ZnO nanostructures: recent progress and new concepts at the borderline. *Polym. Int.* **61**(3), 355–373 (2012)
5. Ponder, J.W., Case, D.A.: Force fields for protein simulations. *Adv. Protein. Chem.* **66**, 27–85 (2003)
6. Brooks, B.R., Brooks III, C.L., Mackerell, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A.R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R.W., Post, C.B., Pu, J.Z., Schaefer, M., Tidor, B., Venable, R.M., Woodcock, H.L., Wu, X., Yang, W., York, D.M., Karplus, M.: CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **30**(10), 1545–1615 (2009)
7. Jorgensen, W.L., Tirado-Rives, J.: The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110**(6), 1657–1666 (1988)
8. Schmid, N., Eichenberger, A.P., Choutko, A., Riniker, S., Winger, M., Mark, A.E., van Gunsteren, W.F.: Definition and testing of the GROMOS force-field versions 54a7 and 54b7. *Eur. Biophys. J.* **40**, 843–856 (2011)
9. Marrink, S.J., Risselada, H.J., Yefimov, S., Tieleman, D.P., de Vries, A.H.: The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **111**(27), 7812–7824 (2007)
10. Cieplak, P., Dupradeau, F.-Y., Duan, Y., Wang, J.: Polarization effects in molecular mechanical force fields. *J. Phys. Condens. Matter* **21**(33), 333102 (2009)
11. Niketić, S.R., Rasmussen, K.: The Consistent Force Field: A Documentation. Lecture Notes in Chemistry, vol. 3. Springer, Berlin/Heidelberg (1977)
12. Rappe, A.K., Casewit, C.J., Colwell, K.S., Goddard III, W.A., Skiff, W.M.: UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**(25), 10024–10035 (1992)
13. Allinger, N.L., Chen, K.-H., Lii, J.-H., Durkin, K.A.: Alcohols, ethers, carbohydrates, and related compounds. I. The MM4 force field for simple compounds. *J. Comput. Chem.* **24**(12), 1447–1472 (2003)
14. Sun, H., Ren, P., Fried, J.R.: The COMPASS force field: parameterization and validation for phosphazenes. *Comput. Theor. Polym. Sci.* **8**(1–2), 229–246 (1998)
15. Ponder, J.W., Wu, C., Ren, P., Pande, V.S., Chodera, J.D., Schnieders, M.J., Haque, I., Mobley, D.L.,

- Lambrecht, D.S., DiStasio Jr., R.A., Head-Gordon, M., Clark, G.N.I., Johnson, M.E., Head-Gordon, T.: Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B* **114**(8), 2549–2564 (2010)
16. Matsui, M., Akaogi, M.: Molecular dynamics simulation of the structural and physical properties of the four polymorphs of TiO₂. *Mol. Simul.* **6**(4–6), 239–244 (1991)
17. Kulkarni, A.J., Zhou, M., Ke, F.J.: Orientation and size dependence of the elastic properties of zinc oxide nanobelts. *Nanotechnology* **16**(12), 2749–2756 (2005)
18. Raymand, D., van Duin, A.C.T., Baudin, M., Hermansson, K.: A reactive force field (ReaxFF) for zinc oxide. *Surf. Sci.* **602**(5), 1020–1031 (2008)
19. Raju, M., Kim, S.-Y., van Duin, A.C.T., Fichthorn, K. A.: ReaxFF reactive force field study of the dissociation of water on titania surfaces. *J. Phys. Chem. A C* **117**(20), 10558–10572 (2013)
20. Woodcock, L.V., Angell, C.A., Cheeseman, P.: Molecular dynamics studies of the vitreous state: simple ionic systems and silica. *J. Chem. Phys.* **65**(4), 1565–1577 (1976)
21. Feuston, B.P., Garofalini, S.H.: Empirical three-body potential for vitreous silica. *J. Chem. Phys.* **89**(9), 5818–5824 (1988)
22. Vessal, B., Amini, M., Fincham, D., Catlow, C.R.A.: Water-like melting behavior of SiO₂ investigated by the molecular dynamics simulation technique. *Philos. Mag. B* **60**(6), 753–775 (1989)
23. van Beest, B.W.H., Kramer, G.J., van Santen, R.A.: Force fields for silicas and aluminophosphates based on ab initio calculations. *Phys. Rev. Lett.* **64**(16), 1955–1958 (1990)
24. Huff, N.T., Demiralp, E., Çagin, T., Goddard III, W.A.: Factors affecting molecular dynamics simulated vitreous silica structures. *J. Non Cryst. Solids* **253**(1–3), 133–142 (1999)
25. Takada, A., Richet, P., Catlow, C.R.A., Price, G.D.: Molecular dynamics simulations of vitreous silica structures. *J. Non Cryst. Solids* **345–346**, 224–229 (2004)
26. van Duin, A.C.T., Strachan, A., Stewman, S., Zhang, Q., Xu, X., Goddard III, W.A.: ReaxFF_{SiO} reactive force field for silicon and silicon oxide systems. *J. Phys. Chem. A* **107**(19), 3803–3811 (2003)
27. Saba, M.I., Calzia, V., Melis, C., Colombo, L., Mattoni, A.: Atomistic investigation of the solid-liquid interface between the crystalline zinc oxide surface and the liquid tetrahydrofuran solvent. *J. Phys. Chem. C* **116**(23), 12644–12648 (2012)
28. Mallochi, G., Binda, M., Petrozza, A., Mattoni, A.: Role of molecular thermodynamical processes at functionalized polymer/metaloxide interfaces for photovoltaics. *J. Phys. Chem. C* **117**(27), 13894–13901 (2013)
29. Melis, C., Raiteri, P., Colombo, L., Mattoni, A.: Self-assembling of Zinc Phthalocyanines on ZnO (10̄10) surface through multiple time scales. *ACS Nano* **5**(12), 9639–9647 (2011)
30. Allen, M.P., Tildesley, D.J.: Computer Simulation of Liquids. Clarendon, Oxford (1989)
31. Kang, Y., Li, X., Tu, Y., Wang, Q., Ågren, H.: On the mechanism of protein adsorption onto hydroxylated and nonhydroxylated TiO₂ surfaces. *J. Phys. Chem. C* **114**(34), 14496–14502 (2010)
32. Ham, H.O., Park, S.H., Kurutz, J.W., Szleifer, I.G., Messersmith, P.B.: Antifouling glycocalyx-mimetic peptoids. *J. Am. Chem. Soc.* **135**(35), 13015–13022 (2013)
33. Kenausis, G.L., Vörös, J., Elbert, D.L., Huang, N., Hofer, R., Ruiz-Taylor, L., Textor, M., Hubbell, J.A., Spencer, N.D.: Poly(l-lysine)-g-poly(ethylene glycol) layers on metal oxide surfaces: attachment mechanism and effects of polymer architecture on resistance to protein adsorption. *J. Phys. Chem. B* **104**(14), 3298–3309 (2000)
34. Saba, M.I., Mattoni, A.: Effect of thermodynamics and curvature on the crystallinity of P3HT thin films on ZnO: insights from atomistic simulations. *J. Phys. Chem. C* **118**(9), 4687–4694 (2014)
35. Caddeo, C., Dessì, R., Melis, C., Colombo, L., Mattoni, A.: Poly(3-hexylthiophene) adhesion on zinc oxide nanoneedles. *J. Phys. Chem. C* **115**(34), 16833–16837 (2011)
36. Meredig, B., Salleo, A., Gee, R.: Ordering of poly (3-hexylthiophene) nanocrystallites on the basis of substrate surface energy. *ACS Nano* **3**(10), 2881–2886 (2009)
37. Mattioli, G., Dkhil, S.B., Saba, M.I., Mallochi, G., Melis, C., Alippi, P., Filippone, F., Giannozzi, P., Thakur, A.K., Gaceur, M., Margeat, O., Diallo, A.K., Videlot-Ackermann, C., Ackermann, J., Bonapasta, A. A., Mattoni, A.: Interfacial engineering of P3HT/ZnO hybrid solar cells using phthalocyanines: a joint theoretical and experimental investigation. *Adv. Energy Mater.* **4**, 1301694 (2014)

Simulations of Solid Interfaces

► Surface Modeling of Ceramic Biomaterials

S

Single Cell Analysis

► Electrical Impedance Tomography for Single-Cell Imaging

Single Cell Impedance Spectroscopy

► Electrical Impedance Cytometry

Single-Cell Electrical Impedance Spectroscopy

► Single-Cell Impedance Spectroscopy

Single-Cell Impedance Spectroscopy

Jian Chen¹, Nika Shakiba², Qingyuan Tan² and Yu Sun³

¹State Key Laboratory of Transducer Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing, People's Republic of China

²Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada

³Department of Mechanical and Industrial Engineering and Institute of Biomaterials and Biomedical Engineering and Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada

Synonyms

Impedance measurement of single cells; Impedance spectroscopy for single-cell analysis; Single-cell electrical impedance spectroscopy

Definition

Single-cell impedance spectroscopy is a technique that operates by applying a frequency-dependent excitation signal on a single cell positioned in between two measurement microelectrodes. The current response is measured as a function of frequency. By interpreting the impedance profile, dielectric properties of a single cell such as cell membrane capacitance and cytoplasmic resistance are obtained.

Overview

Historical Development

In conventional cell electrical impedance spectroscopy, the impedance of multiple cells as a

whole is measured using an AC excitation signal. The cell suspension is held in a container with two electrodes. Current is measured as a function of frequency to determine the electrical properties of the cells in the suspension.

Recent advances in microfabrication and lab-on-a-chip technologies enable the development of electrical impedance spectroscopic devices to measure impedance profiles of cells at the single-cell level, providing useful biophysical characteristics of single cells and promising potentially noninvasive, label-free approaches for analyzing cells.

Working Principle

Single-cell impedance spectroscopy measures impedance profiles of a cell positioned in between two microelectrodes [1]. A low-magnitude AC voltage, $\tilde{U}(j\omega)$, over a range of frequencies is used as the excitation signal. The current response, $\tilde{I}(j\omega)$, is measured. The impedance $\tilde{Z}(j\omega)$ is

$$\tilde{Z}(j\omega) = \frac{\tilde{U}(j\omega)}{\tilde{I}(j\omega)} = \tilde{Z}_{RE} + j\tilde{Z}_{IM} \quad (1)$$

where \tilde{Z}_{RE} and \tilde{Z}_{IM} are the real and imaginary parts of impedance. The real part is termed resistance while the imaginary part is termed reactance. The magnitude and phase angle are

$$|\tilde{Z}| = \sqrt{\tilde{Z}_{RE}^2 + \tilde{Z}_{IM}^2} \quad (2)$$

and

$$\angle \tilde{Z} = \arctan \left(\frac{\tilde{Z}_{IM}}{\tilde{Z}_{RE}} \right) \quad (3)$$

Advantages and Applications

Compared to impedance measurement on a cell suspension, single-cell impedance spectroscopy interrogates the property of a single cell (vs. a cell population). Theoretical analysis can also become simpler because it is not necessary to take into account electrical interactions among cells [2].

The technique of single-cell impedance spectroscopy has been used as a noninvasive method to quantify the physiological state of single cells. As a biophysical marker, single-cell impedance profiles have also been utilized to distinguish normal cells from abnormal cells (e.g., human cancer cells with different metastatic potential and malaria-infected red blood cells) [3, 4].

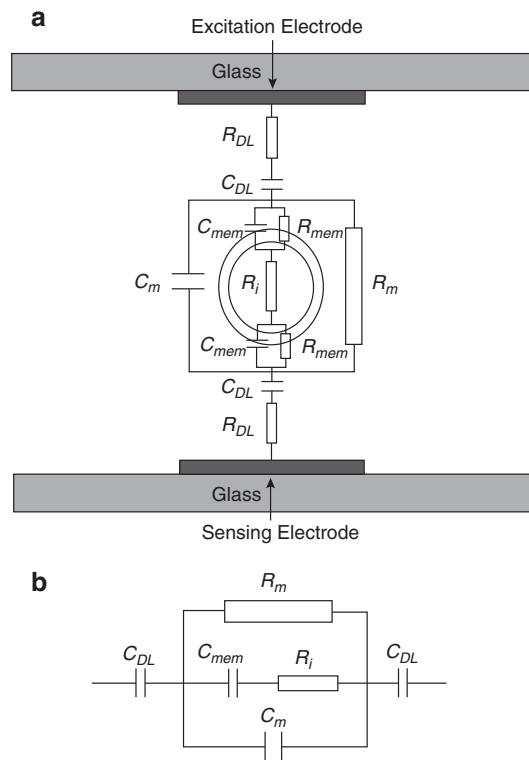
Methodology

Equivalent Circuit Model

Figure 1a [1] shows an equivalent circuit for interpreting single-cell impedance measurement data [1]. In this circuit model, the cell is represented by a capacitor in series with a resistor, with elements connected in parallel with the capacitor and resistor (i.e., the cell) representing the suspending medium. At low frequencies, the thin cell membrane gives rise to a large capacitance. As the frequency increases, the reactive component of this element tends to zero out, and the cell internal properties are represented by the resistor. In this circuit model, the membrane is assumed to have a high resistance and the cytoplasmic capacitance is ignored. All the circuit elements are functions of the volume fraction or cell size.

The electrical double layer has an influence on measurements at low frequencies (below 1 MHz for high-conductivity solutions). This is generally modeled as a constant phase angle, represented by a resistor (R_{DL}) and capacitor (C_{DL}) in series. As shown in Fig. 1a, the double layer is in series with the network model. In the simplified model shown in Fig. 1b, the double layer is assumed to be purely capacitive (C_{DL}), with a value given by the product of the inverse Debye length and the permittivity of the medium.

The simplified circuit in Fig. 1b [1] shows that at very low frequencies current flow is blocked by the double layer capacitor, and only the impedance of the double layer is measured. As the measurement frequency increases, this capacitor is gradually short-circuited and the excitation voltage charges the cell in suspension. It takes a finite amount of time



Single-Cell Impedance Spectroscopy, Fig. 1 (a) The equivalent circuit model for single-cell impedance spectroscopy. R_{DL} and C_{DL} represent the resistance and capacitance of the electrical double layer, R_m and C_m the resistance and capacitance of the medium, R_{mem} and C_{mem} the resistance and capacitance of the cell membrane and R_i the resistance of the cytoplasm. (b) The simplified equivalent circuit model neglecting the electrical double layer resistance and the membrane resistance (Reproduced with permission from Ref. [1])

to charge the membrane through the extracellular and intracellular fluid, resulting in two dielectric dispersions in the frequency range of 1–100 MHz. The lower frequency dispersion is governed by the polarization of the cell membrane. Measurement of this parameter provides information about the dielectric properties of the membrane. The higher-frequency dispersion is governed by the polarization of the cytoplasm and the suspending medium as the membrane is short-circuited at these frequencies. This second dispersion is generally small and difficult to measure using impedance spectroscopy techniques.

With the effect of the double layer taken into account, the total impedance of the circuit is

$$\begin{aligned}\tilde{Z}(j\omega) = & \frac{1}{j\omega C_{DL}} \\ & + \frac{R_m(1+j\omega R_i C_{mem})}{j\omega R_m C_{mem} + (1+j\omega R_i C_{mem})(1+j\omega R_m C_m)}\end{aligned}\quad (4)$$

Single-Cell Impedance Spectroscopy on Stationary or Moving Cells

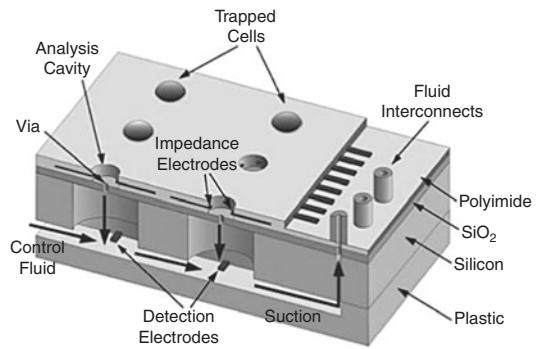
Single-cell impedance spectroscopy in microfluidic devices can be divided into two main categories: measurements on either stationary or moving cells. For measuring a stationary cell, a cell can be positioned/trapped in between microelectrodes using several approaches. Cells can also be controlled to pass the measurement electrodes at the speed of microfluidic flow.

1. Cell aspiration [5]

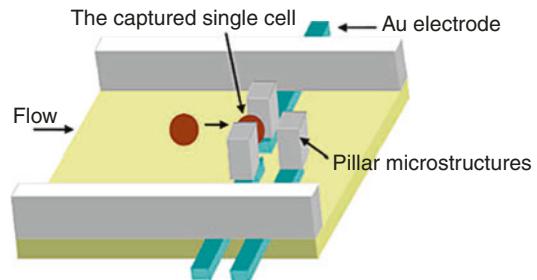
- (a) Working mechanism: In the aspiration approach, a negative pressure is applied to trap a single cell in a channel opening. Measurement electrodes can be built into such channel openings. A trapped cell is forced by the applied negative pressure to contact the electrodes (Fig. 2) [5]. The impedance profile of the cell is then obtained.
- (b) Advantages: The main advantage of the aspiration approach for capturing a cell is the seal formed between the measurement electrodes and the cell. The suction force is sufficient to hold the cell in place throughout the measurement process, especially when the cell-capturing channel has a more or less circular cross section.
- (c) Potential limitations: It has been suggested that mechanical deformations of a cell due to the suction force may lead to changes in the electrical properties of the cell, thus, affecting the impedance profile measured.

2. Hydrodynamic trapping [6]

- (a) Working mechanism: In hydrodynamic trapping, cells are trapped by



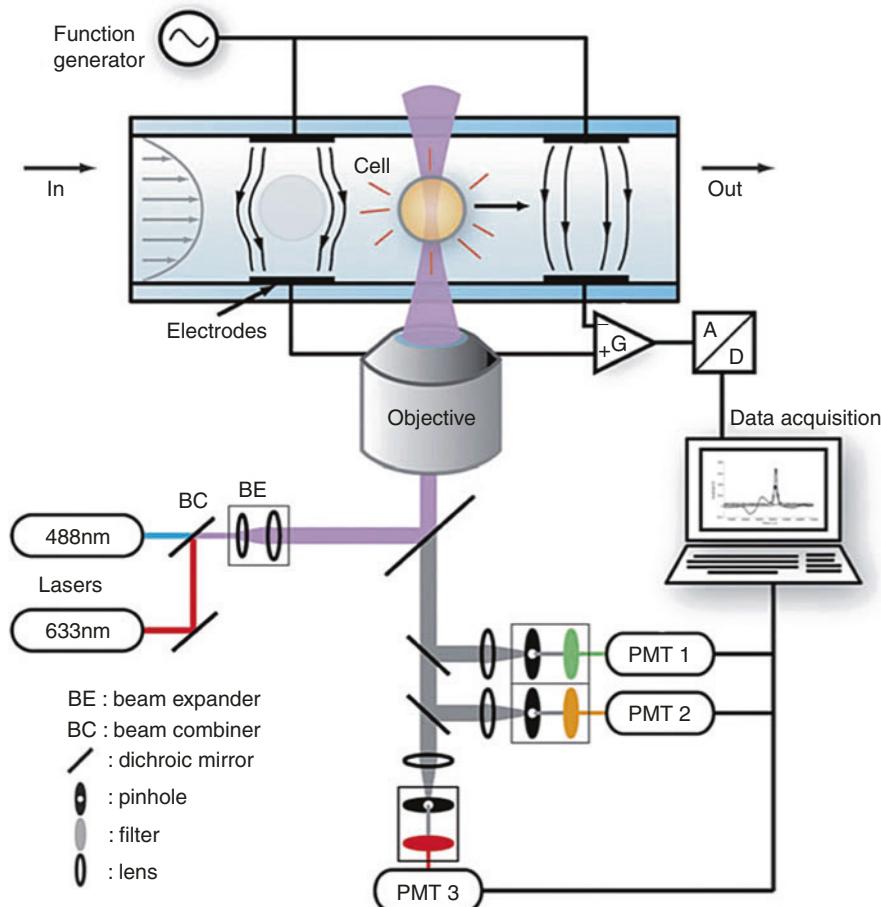
Single-Cell Impedance Spectroscopy, Fig. 2 Schematic view of an aspiration-based single-cell impedance spectroscopy system. The system is composed of an array of analysis cavities, each capable of analyzing a single cell at a time, each containing a mechanical fluid via with flow control to hold the cell in position during the impedance analysis operation, and multiple electrodes surrounding the circumference of the cell to measure the electrical characteristics of the captured cell (Reproduced with permission from Ref. [5])



Single-Cell Impedance Spectroscopy, Fig. 3 Schematic view of a microfluidic device for single-cell impedance spectroscopy using the concept of hydrodynamic trapping for single-cell immobilization (Reproduced with permission from Ref. [6])

microfabricated weirs, used as filters on top of electrodes. By adjusting flow rates, cells can be captured on measurement electrodes for impedance data recording (Fig. 3) [6].

- (b) Advantages: The main advantage of this approach is the feasibility to realize large-array single-cell trapping and potentially high-throughput impedance measurement.
- (c) Potential limitations: Contact between the trapped cell and the measurement



Single-Cell Impedance Spectroscopy, Fig. 4 Schematic view of differential impedance micro flow cytometry. Each cell's impedance signal is recorded by a differential pair of microelectrodes using media without a

single cell passing as a reference. In this setup, the effect of the electrical double layer on impedance profiles is canceled (Reproduced with permission from Ref. [7])

microelectrodes underneath can be problematic since there is no force to hold the cell in close contact with the microelectrodes.

3. Flow cytometry [7]

(a) Working mechanism: Flow cytometry is a technique that allows for the analysis of cells in suspension with single-cell precision. A laminar flow carries suspended cells through the measurement location. Each cell's impedance signal is recorded by a differential pair of microelectrodes, using the surrounding media without a cell as reference. Microfabricated devices and electronic circuits allow simultaneous

impedance measurements at multiple frequencies (Fig. 4) [7].

- (b) Advantages: Flow cytometry is advantageous in that it allows for rapid analysis of cells in suspension with a high throughput. Furthermore, flow cytometry can also allow for the sorting of cells into subpopulations based on their impedance measurement profiles.
- (c) Potential limitations: There is no contact between the cell and measurement electrodes and thus, the impedimentary contribution of the current leakage in the medium around the cell is significant.

Key Research Findings

Effect of Electrical Double Layer on Impedance Profile

When a polarizable electrode (i.e., an electrode operated at a regime where no charge transfer reaction occurs at the surface) comes into contact with an electrolyte, charges from the electrolyte opposite in sign to the charges present on the surface of the electrode move the electrode/electrolyte interface and provide a localized condition of electroneutrality as well as a layer of charge, termed the electrical double layer.

The electrical double layer has an influence on cell impedance measurements at low frequencies, which is generally represented as a capacitor. The most effective way to minimize the electrical double layer is to maximize the electrolyte-electrode interface area. This interface area enhancement can be achieved either by mechanically roughing the electrode surface to an electrode-electrolyte interface area that is effectively larger than the actual electrode surface, or to use chemical treatments that lead to a high electrode-electrolyte interface area.

Comparison of Impedance Measurement Mechanisms

The essential feature of measurements on a stationary cell trapped/positioned between electrodes is a tight seal between the cell and electrodes. In this configuration, the impedimentary contribution of the cell shunt (i.e., current leakage in the medium around the cell) is insignificant.

The advantage of impedance cytometry is differential impedance measurement that minimizes the effect of the electrical double layer. In an alternating fashion, each measurement electrode-ground electrode pair without a cell serves as a reference to the other pair, over which a cell passes.

Future Work

A single-cell impedance spectroscopy measurement system ideally should have a high testing

throughput and the capability of sorting single cells based on impedance measurement results. The potential combination of single-cell impedance spectroscopy with other detection/measurement methods, such as fluorescent detection and the use of biochemical markers, may prove powerful for a range of applications, such as disease diagnostics and rare cell isolation.

References

1. Morgan, H., Sun, T., Holmes, D., Gawad, S., Green, N.: Single cell dielectric spectroscopy. *J. Phys. D Appl. Phys.* **40**, 61–70 (2007)
2. Valero, A., Braschler, T., Renaud, P.: A unified approach to dielectric single cell analysis: impedance and dielectrophoretic force spectroscopy. *Lab Chip* **10**, 2216–2225 (2010)
3. Sun, T., Morgan, H.: Single-cell microfluidic impedance cytometry: a review. *Microfluid. Nanofluid.* **8**, 423–443 (2010)
4. Cheung, K.C., Di Berardino, M., Schade-Kampmann, G., Hebeisen, M., Pierzchalski, A., Bocsi, J., Mittag, A., Tarnok, A.: Microfluidic impedance-based flow cytometry. *Cytometry A* **77A**, 648–666 (2010)
5. Han, K.H., Han, A., Frazer, A.B.: Microsystems for isolation and electrophysiological analysis of breast cancer cells from blood. *Biosens. Bioelectron.* **21**, 1907–1914 (2006)
6. Jang, L.S., Wang, M.H.: Microfluidic device for cell capture and impedance measurement. *Biomed. Microdevices* **9**, 737–743 (2007)
7. Holmes, D., Pettigrew, D., Reccius, C.H., Gwyer, J.D., Berkel, C.V., Holloway, J., Daviesb, D.E., Morgan, H.: Leukocyte analysis and differentiation using high speed microfluidic single cell impedance cytometry. *Lab Chip* **9**, 2881–2889 (2009)

Single-Walled Carbon Nanotubes (SWCNTs)

► Chemical Vapor Deposition (CVD)

Sintered

► Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm

siRNA Delivery

► RNAi in Biomedicine and Drug Delivery

Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures

Julia R. Greer

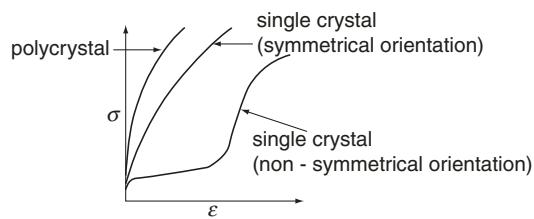
Division of Engineering and Applied Sciences,
California Institute of Technology, Pasadena,
CA, USA

Definition

“Size-dependent plasticity” here refers to the strength of metallic samples being a strong function of their size when their dimensions are reduced to the micron and below scales. The notion of reduced sample size applies to all three dimensions, i.e., stand-alone, or one-dimensional (1D) nano and microstructures rather than thin films (2D), where only their thicknesses are reduced to nano- and micro-scales, or to the small-scale deformation volumes within a bulk matrix (3D) as would be the case during nanoindentation experiments, for example.

Introduction and Overview of Stress Vs. Strain for Bulk Metals

Pure bulk single crystalline metals like Cu, Ni, Mo, Ti, etc. generally exhibit a convex, continuous stress-strain relationship upon uniaxial deformation – tension or compression – as can be found in any classic “Mechanical Behavior of Materials” book and as is also schematically shown in Fig. 1. For the low-symmetry orientations, i.e., where only a single slip system is activated, the deformation immediately following yield is called the “easy glide” region, which is characterized by a low hardening rate. Such low hardening rate stems from the dislocations traveling large distances unimpeded in their glide



Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures, Fig. 1 Schematic of typical stress–strain curves for bulk metals: comparing polycrystalline and single crystalline (in high- and low-symmetry orientations) microstructures of the same metal

planes as only one set of crystallographic slip planes is active, and the dislocations are not forced to overcome closely spaced obstacles such as impurities or other dislocations in the course of their motion. A simple way to think of it is the following: Each dislocation travels a distance L before encountering an obstacle, which pins it. The shear strain associated with this motion is: $d\gamma = bLdp$, where b is the Burgers vector, L is the dislocation mean free path, and p is the mobile dislocation density [1]. The distance traveled by each dislocation then scales with $1/\sqrt{p}$, i.e., $L = \beta/\sqrt{p}$, where β may be on the order of 1,000 since the dislocations traveling in parallel crystallographic planes have very limited interactions. The hardening corresponding to this “easy glide” region is a result of dislocation storage through the well-known Taylor relation: $\tau = \alpha pb\sqrt{p}$, where $\alpha \sim 0.2$, which results in a very low hardening rate $d\tau/d\gamma_{low-symmetry} = 10^{-4}\mu$. In high-symmetry orientations, multiple symmetric slip systems are activated, and dislocations come into close encounters with one another as they are traveling toward one another in the symmetric slip planes, resulting in a pronounced dislocation density increase, which in turn, drives the high rate of hardening: $d\tau/d\gamma_{high-symmetry} = 0.01\mu$. This hardening in bulk single crystals stems from dislocation multiplication processes arising from the well-established dislocation interactions, which are followed by the dislocation networks formation processes. Therefore, in bulk single crystals – regardless of the sample size – the

dislocations multiply in the course of deformation, thereby creating dense networks and dislocation sub-structures, which require the application of higher stresses to move the additional gliding dislocations through these obstacles, thereby carrying out plastic deformation. This holds true for crystals of any size with greater-than-several-microns dimensions. As a result, bulk samples of the same single crystalline material also exhibit identical yield strengths, flow stresses, and hardening rates.

Emergence of Size Effects in Single Crystals at the Nanoscale

In a striking deviation from this classical depiction, in the last 5–6 years, it was ubiquitously demonstrated that at the micron- and sub-micron scales, the sample size dramatically affects crystalline strength – even in the absence of any constraining effects, strain gradients, and grain boundaries – as revealed by room-temperature uniaxial compression experiments on a wide range of single crystalline metallic nano-pillars (for reviews, see [2–4]). In these studies, cylindrical nano-pillars were fabricated mainly by the use of the focused ion beam (FIB), as well as with some FIB-less methodologies, and remarkably, the results of these reports for all metallic single crystals show power law dependence between the flow stress and sample size, regardless of fabrication technique, of the form $\sigma = C \times D^{-k}$ with the exponent k strongly dependent on the initial crystal structure (i.e., face-centered cubic (fcc), body-centered cubic (bcc), hexagonal close-packed (hcp), etc.), experimental aspects (i.e., sample aspect ratio, lateral stiffness of the instrument, etc.), and dislocation density. A particularly auspicious example is the unique scaling of strength with pillar diameter with the exponent of ~ -0.6 exhibited by nearly all non-pristine (i.e., containing initial dislocations) fcc nano-pillars subjected to uniaxial compression or tension, as illustrated in Fig. 2.

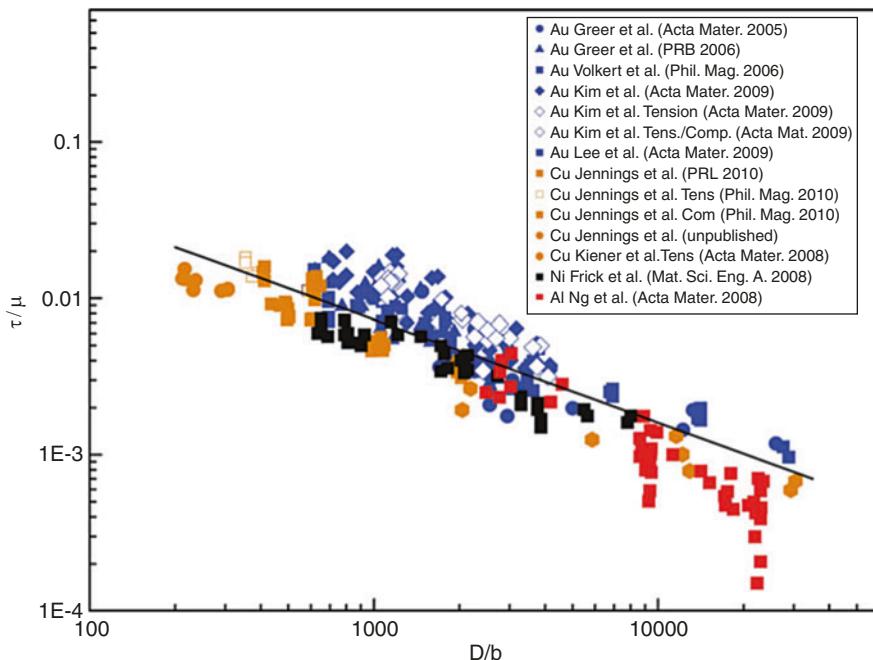
The notion of “size effects” applied specifically to the strength of single crystalline metals dates back to the works of Brenner [5, 6], who

demonstrated that dislocation-free Cu and Ag whiskers attained very high strengths upon uniaxial tension. Although significant efforts have since been dedicated to studying single crystalline plasticity in confined dimensions, until ~5 years ago, most of these research thrusts were focused on thin films, whose yield strengths increased with decreasing film thickness (for example, [7]). The renaissance of small-scale mechanical testing on free-standing vertical pillars is largely due to the original work of Uchic et al. who reported higher compressive strengths attained by focused ion beam (FIB)-machined cylindrical Ni micro-pillars [8]. Greer and Nix then extended this robust and elegant methodology into the nanoscale regime, where $<001>$ -oriented Au nano-pillars with diameters below 1 μm were reported to be 50 times higher than bulk [9], and today numerous groups are pursuing this type of uniaxial nanomechanical testing of materials ranging from single crystalline metals to lithiated battery anodic materials, ceramics, irradiated materials, shape memory alloys, nano-twinned and nanocrystalline metals, metallic glasses, superalloys, and nanolaminates. The results of many of these studies are overviewed in detail in three existing reviews: [2–4]. Figure 3 shows representative images of before- and after-testing single crystalline Nb nano-pillars, as well as the representative stress-strain curves with clear discrete characteristics and size effect [10]. Further adapting this methodology, the exploration of size effects in plasticity for a large variety of materials ensued, albeit mainly focusing on fcc structures. More recently, investigating plasticity in small volumes has been further advanced through uniaxial tensile experiments, usually conducted inside of in-situ mechanical deformation instruments custom built by some research groups [9–12].

Key Research Findings

Experimental Findings

To date, uniaxial compression and tension tests have been performed on Ni and Ni-based superalloys, Au, Cu, and Al (as-fabricated and intentionally passivated) [2–4]. Beyond single



Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures, Fig. 2 Resolved shear strength normalized by the shear modulus as a function of diameter (normalized by the Burgers vector) of

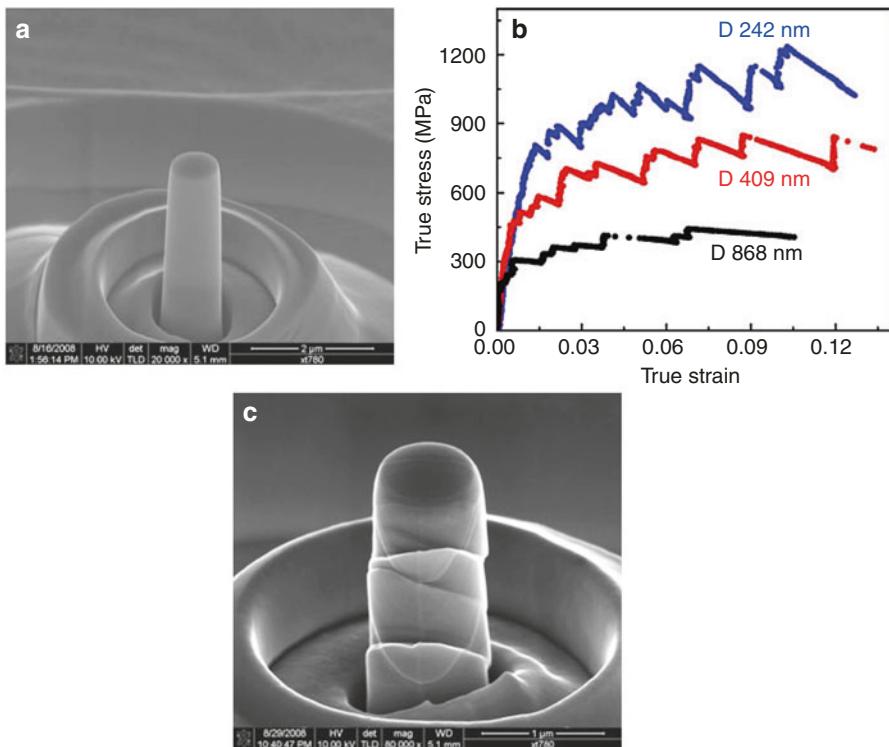
experimentally determined compressive strengths of most face-centered cubic metallic nano-pillars published to date (Reprinted with permission from Ref. [2])

crystalline fcc metals, mechanical behavior of other single crystals has been published to date: bcc metals (W, Nb, Ta, Mo, and V), hcp metals (Mg, Ti-Al alloy, and Ti), tetragonal metals (In) [2–4]. The reports on nearly all single crystalline metals with non-zero initial dislocation densities (i.e., not whiskers or nano-wires) unanimously demonstrate that their strengths significantly rise with reduced size, with ~ 100 nm-diameter samples sometimes attaining a flow stress $\sim 10\times$ higher than bulk. Intriguingly, unlike in bulk, where the dislocation multiplication processes result in Taylor hardening (introduced in section “[Introduction and Overview of Stress Vs. Strain for Bulk Metals](#)”), the flow stresses in small structures do not appear to scale with the dislocation density. Rather, the global dislocation density appears to decrease upon mechanical loading, while the applied stress required to deform the structure increases (see, for example [11]). Consistent with this “upside down” behavior in the nano- and micron-sized crystals, whereby samples

containing fewer mobile dislocations are stronger than their bulk counterparts, it has been shown that introduction of additional dislocations into the structure (for example, via pre-straining) actually *weakens* these crystals [12, 13]. These findings are antipodal to classical plasticity, where dislocation interactions lead to multiplication, causing higher dislocation densities and requiring higher applied stresses for deformation to continue, as described in the section “[Introduction and Overview of Stress Vs. Strain for Bulk Metals](#)”. The initial dislocation density indeed plays a critical role in the onset of the size effect, as has now been shown by several research groups [2–4].

Computational Findings

Several models attempting to explain the origins of size-dependent flow stress in the absence of strong strain gradients, as well as the stochastic nature of deformation, have been put forth. There are generally three classes of these models: (1) phenomenological theories, which attempt to



Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures, Fig. 3 SEM images of (a) single crystalline Nb nano-pillar as fabricated by the focused ion beam, (c) same pillar after compression showing pronounced crystallographic slip lines. (b) Stress

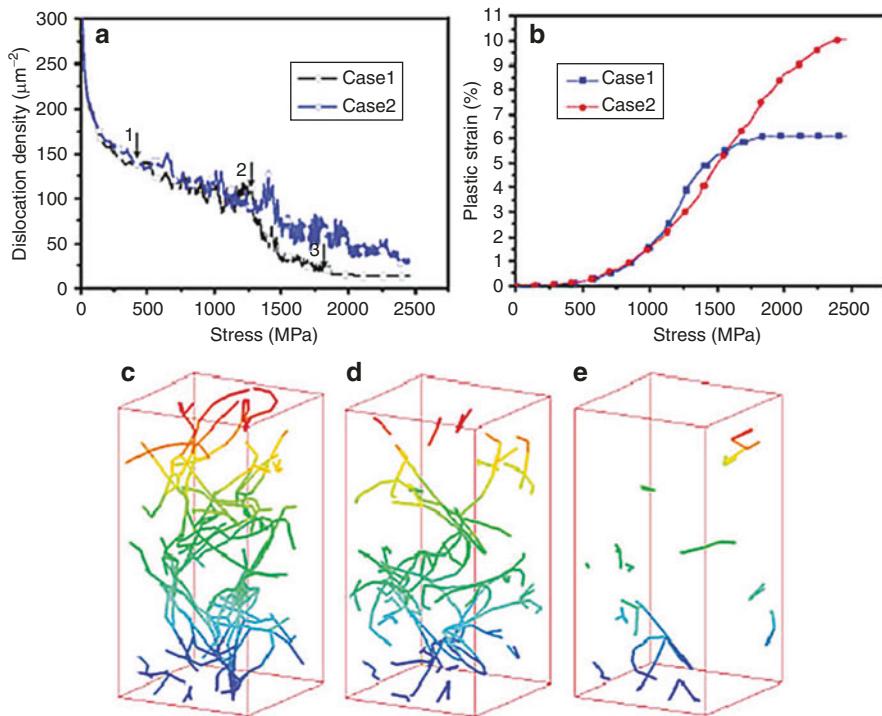
versus strain curves clearly revealing the size effect (smaller is stronger) and the stochastic signature of small-scale deformation (Reprinted with permission from Ref. [10])

describe the physical processes occurring in a nano-single-crystal upon deformation, (2) discrete dislocation dynamics (two- and three-dimensional) simulations (DD), which rely on the entered initial dislocation density and distribution, as well as on the dislocation mobility laws, and (3) molecular dynamics (MD) simulations, which are generally limited to small (50 nm and below) scales and unrealistically high strain rates ($\sim 10^{-8} \text{ s}^{-1}$). An example of (1) is the “hardening by dislocation starvation” theory which hypothesizes that the mobile dislocations inside a small nano-pillar have a greater probability of annihilating at a free surface than of interacting with one another, thereby shifting plasticity into nucleation-controlled regime [14–16]. Other models include source exhaustion hardening, source truncation, and weakest link theory, all of whose general premise involves representing

dislocation source operations in a discrete fashion, and then evaluating the effect of sample size on the source lengths, and therefore on the strengths of their operation [17]. These models capture the ubiquitously observed stochastic signature of the experiments results and show either marginal dislocation storage or no storage at all. An example of an evolved microstructure as revealed by dislocation dynamics simulations is shown in Fig. 4 and clearly shows that upon deformation, the dislocation density decreases, rendering multiplication processes unlikely and shifting plasticity into dislocation nucleation-controlled regime.

Discussion: What Causes the Size effect?

Despite the definition of “ultimate tensile strength” according to Wikipedia stating that



Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures, Fig. 4 Variations in (a) dislocation density and (b) plastic strain with increasing stress. A dislocation avalanche occurs during the second stage.

(c–e) are the microstructures corresponding to points 1, 2, and 3 in (a), respectively (Reprinted with permission from Ref. [20])

UTS is an intensive property and “therefore its value does not depend on the size of the test specimen,” there is now a large body of work that convincingly demonstrates that this does not hold true in the micron and sub-micron size regime. The ubiquitously reported size-dependent strengths in small-scale single crystalline micro- and nano-pillars show the “smaller is stronger” phenomenon as revealed by the existence of a power law dependence between the attained flow stress (σ_{flow}) and pillar diameter (D): $\sigma_{\text{flow}} = C \times D^{-n}$, where n varies with the specific crystal structure, initial dislocation density, and deformation type. Intriguingly, it appears that the power law slope for all fcc metals is unified, on the order of -0.6 , as shown in Fig. 2. This is not the case for body-centered cubic metals, whose power law slopes vary within this crystal structure family, possibly due to the unique potential energy landscape of each bcc metal that a gliding screw dislocation has to overcome. The power law

slopes of bcc metals appear to correlate with the intrinsic lattice resistance of each metal, a.k. a. the Peierls barrier, and can be separated into groups according to the low, mid- and high Peierls barrier: low-barrier group containing Nb (-0.93 , -0.48 , -1.07) and V (-0.79); mid-barrier group containing Mo (-0.44 , -0.38) and Ta (-0.43 , -0.41); and high-barrier group containing W (-0.21 , -0.44) [2–4]. Beyond these two cubic crystalline structures, other types of crystals were also found to exhibit size-dependent strengths, with smaller generally being stronger. For example, hexagonal close-packed (hcp) metals, Mg and Ti, oriented for slip (as opposed to twinning) were also characterized by power law size effects: [11–21] Ti oriented for prismatic slip has the critical resolved shear stress (CRSS) scale inversely with the diameter, resulting in the slope of -1 , while the [3–964] orientation of Mg (basal slip) has the slope of -0.64 [2].

There is currently no physics-based theory that captures these size effects as a function of metal, size, and initial microstructure. Several semi-empirical theories have been proposed, and nearly all identify the initial dislocation density as a key factor in defining the natural microstructural length scale, which in turn, drives the size effect. In an attempt to explain the emergence of higher strengths in small-scale crystals, multiple deformation mechanisms have been proposed, with two distinct ones: (1) single-arm source (SAS) theory first developed by Parthasarathy and Rao et al. [17] applicable to micron-sized pillars, and (2) hardening by dislocation starvation proposed by Greer and Nix [9, 14] for the much smaller, nano-sized pillars. In the SAS theory, the creation of dislocations occurs by the operation of truncated Frank-Read spiral sources, a.k.a. single-arm sources (SAS), whose strength, τ_s , is inversely proportional to their average length, λ , through $\tau_s = k_s \mu \frac{\ln(\bar{\lambda}/b)}{\bar{\lambda}/b}$, where, k_s is a source-hardening constant [17]. The smaller micron-sized pillars are not capable of accommodating large single-arm sources, and therefore require the application of higher stresses to activate the stronger, “shorter” sources, giving rise to higher strengths in smaller pillars.

In the dislocation starvation theory, applicable to the much smaller pillars, with diameters deeply in the sub-micron regime, new dislocations are created via surface nucleation rather than through a multiplicative process of SAS operation as is the case in the micron-sized samples. The surface nucleation gets activated only after a significant fraction of the pre-existing gliding dislocations within the pillar has exited through the free pillar surface, i.e., “starving” the crystal of the necessary mobile dislocations to carry plastic strain. A necessary condition for dislocation starvation is that the largest distance that any gliding dislocation has to travel is smaller than the so-called breeding length, or the distance before it replicates itself, as defined by Gilman [18]. This is likely to be the case in the samples with diameters deeply in the submicrometer range since Ag, for example, has the breeding distance on the order of 0.7 μm . The nucleation stress for a surface source

may be represented as $\sigma = \frac{Q^*}{Q} - \frac{k_B T}{Q} \ln \frac{k_B T N v_0}{E \epsilon Q}$, where the first term corresponds to the athermal contribution and the last term represents thermally activated nature of such events [19]. The activation of surface sources has a significant thermal activation component, as is evident from the $T \ln T$ and strain rate ($\dot{\epsilon}$) dependence of the nucleation stress. Recent experimental studies performed on very small (below 500 nm) Cu nano-pillars are consistent with this surface nucleation model and lay out the single-arm source versus surface source operation regimes as a function of both the pillar size and strain rate [2]. While these theories may appear to be competing, it is likely that they both take place at different pillar sizes: with SAS strengthening occurring in the micron-sized pillars (perhaps, down to ~ 500 nm) and with dislocation starvation followed by surface nucleation prevailing at the smaller sizes, deep in the sub-micron regime.

Summary

- Recent experimental and computational results convincingly demonstrate that the strength of nano- and micron-sized single crystalline metals is indeed size dependent and appears to be well represented by a power law of the form $\sigma = C \times D^{-k}$ with the exponent strongly dependent on the initial microstructure (i.e., dislocation density, lattice resistance, and the presence of impurities).
- With the advent of highly capable nanofabrication and analysis instruments, as well as of the sophisticated computational tools, the society is ever closer to developing a more complete understanding of size-dependent mechanical behavior of small-scale structures. To date, it is generally agreed that non-pristine (i.e., containing dislocations) micron-sized fcc structures attain higher strengths at smaller sizes through dislocation multiplication processes produced by the operation of single-arm dislocation sources, whose lengths scale with the sample size. The operation of

these truncated sources, which become exhausted in the course of deformation, drives the corresponding strength increase in these small-scale structures. On the contrary, when the sample dimensions are further reduced to the nano-sized regime, plasticity likely occurs via dislocation nucleation from surface sources (rather than through a multiplicative process), and the higher strengths arise due to the lower probability of finding weaker dislocation sources in smaller structures.

Acknowledgments JRG gratefully acknowledges the financial support of the National Science Foundation (NSF) CAREER grant (DMR-0748267) and the Office of Naval Research (ONR) Grant No. N00014-09-1-0883. The author is particularly grateful to W.D. Nix, A.T. Jennings, D. Jang, J.-Y. Kim, Q. Sun, A. Ngan, C. Weinberger, J. Li, and D. Gianola for useful discussions.

References

1. Nix, W.D.: MSE 208: Mechanical behavior of materials class notes. (1980)
2. Greer, J.R., de Hosson, J.T.M.: Critical review: plasticity in small-sized metallic systems: intrinsic versus extrinsic size effect. *Prog. Mater. Sci.* **56**, 654–724 (2011)
3. Kraft, O., Gruber, P.A., Monig, R., Weygand, D.: Plasticity in confined dimensions. *Annu. Rev. Mater. Res.* **40**, 293 (2010)
4. Uchic, M.D., Shade, P.A., Dimiduk, D.M.: Plasticity of micrometer-scale single crystals in compression. *Annu. Rev. Mater. Res.* **39**, 361–386 (2009)
5. Brenner, S.S.: Tensile strength of whiskers. *J. Appl. Phys.* **27**, 1484–1490 (1956)
6. Brenner, S.S.: Growth and properties of “whiskers”. *Science* **128**, 568 (1958)
7. Thompson, C.V.: The yield stress of polycrystalline thin films. *J. Mater. Res.* **8**, 237 (1993)
8. Uchic, M.D., Dimiduk, D.M., Florando, J.N., Nix, W. D.: Sample dimensions influence strength and crystal plasticity. *Science* **305**, 986 (2004)
9. Greer, J.R., Oliver, W.C., Nix, W.D.: Size dependence of mechanical properties of gold at the micron scale in the absence of strain gradients. *Acta Mater.* **53**, 1821 (2005)
10. Kim, J.-Y., Jang, D., Greer, J.R.: Insights into deformation behavior and microstructure evolution in Nb single crystalline nano-pillars under uniaxial tension and compression. *Scr. Mater.* **61**, 300 (2009)
11. Tang, H., Schwartz, K.W., Espinosa, H.D.: Dislocation escape-related size effects in single-crystal micropillars under uniaxial compression. *Acta Mater.* **55**, 1607 (2007)
12. Bei, H., Shim, S., Pharr, G.M., George, E.P.: Effects of pre-strain on the compressive stress-strain response of Mo-alloy single-crystal micropillars. *Acta Mater.* **56**, 4762 (2008)
13. Lee, S., Han, S., Nix, W.D.: Uniaxial compression of fcc Au nanopillars on an MgO substrate: the effects of prestraining and annealing. *Acta Mater.* **57**, 4404 (2009)
14. Greer, J.R., Nix, W.D.: Nanoscale gold pillars strengthened through dislocation starvation. *Phys. Rev. B* **73**, 245410 (2006)
15. Shan, Z.W., Mishra, R., Syed, S.A., Warren, O.L., Minor, A.M.: Mechanical annealing and source-limited deformation in submicron-diameter Ni crystals. *Nat. Mater.* **7**, 115–119 (2008)
16. Weinberger, C., Cai, W.: Surface controlled dislocation multiplication in metal micro-pillars. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 14304 (2008)
17. Rao, S.I., Dimiduk, D.M., Parthasarathy, T.A., Uchic, M.D., Tang, M., Woodward, C.: Athermal mechanisms of size-dependent crystal flow gleaned from three-dimensional discrete dislocation simulations. *Acta Mater.* **56**, 3245 (2008)
18. Gilman, J.J.: *Micromechanics of Flow in Solids*. McGraw-Hill, New York (1969)
19. Zhu, T., Li, J., Samanta, A., Leach, A., Gall, K.: Temperature and strain rate dependence of surface dislocation nucleation. *Phys. Rev. Lett.* **100**, 025502 (2008)
20. Liu, Z.L., Liu, X.M., Zhuang, Z., You, X.C.: Atypical three-stage-hardening mechanical behavior of Cu single-crystal micropillars. *Scr. Mater.* **60**, 594 (2009)

Small Angle X-Ray Scattering in Grazing Incidence Geometry

► Selected Synchrotron Radiation Techniques

S

Small Unilamellar Vesicle (SUV)

► Liposomes

Small-Angle Neutron Scattering

► Small-Angle Scattering of Nanostructures and Nanomaterials

Small-Angle Scattering

► [Small-Angle Scattering of Nanostructures and Nanomaterials](#)

Small-Angle Scattering of Nanostructures and Nanomaterials

M. Laver

Laboratory for Neutron Scattering, Paul Scherrer Institut, Villigen, Switzerland

Materials Research Division, Risø DTU, Technical University of Denmark, Roskilde, Denmark

Nano-Science Center, Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

Department of Materials Science and Engineering, University of Maryland, College Park, MD, USA

Synonyms

SANS; SAXS; Small-angle neutron scattering; Small-angle scattering; Small-angle X-ray scattering

Definition

Small-angle scattering of nanostructures and nanomaterials encompasses the measurements of scattering from structures with length scales ranging between the near-atomic (nanometer) to the near-optical (micrometer), using beams of nanometer wavelengths or less.

Overview

Small-angle scattering reveals structural features on length scales between the near-atomic

(nanometer) to the near-optical (micrometer). With such versatility, the technique has made an astounding impact in many fields of research including polymer systems [1, 2], complex fluids [3], biology [4], condensed matter physics, and materials science [1, 5]. The methodology for performing small-angle scattering studies extends back to the 1930s [6], following the first small-angle X-ray scattering (SAXS) studies. Subsequently, synchrotron sources have enabled explorations with ever smaller sample amounts, in complex sample environments. Laboratory SAXS instruments are now routinely used for sample characterization in many research institutions. The creation of large-scale facilities producing neutrons for research has proved to be instrumental for small-angle neutron scattering (SANS) studies. As X-rays and neutrons interact differently with matter, SAXS and SANS are quintessentially complementary, and a particular technique may frequently become an indispensable probe for a particular application.

This entry is organized as follows: in section “[Key Principles](#)”, the general concepts of small-angle scattering are explained. In section “[X-Rays or Neutrons](#)”, the advantages and practical aspects of X-ray and neutron techniques are compared. In section “[Magnetic Neutron Scattering](#)”, the discussion is focused on small-angle neutron scattering from magnetic structures, and neutron polarization analysis (section “[Neutron Polarization Analysis](#)”) is provided as an example of a developing field for small-angle studies. The concluding section (section “[Summary and Outlook](#)”) incorporates an outlook toward novel directions for the small-angle scattering technique.

An overview of symbols used repeatedly throughout this entry is provided in Table 1.

Key Principles

Bragg Diffraction

Small-angle scattering patterns are frequently continuous in nature, rather than consisting

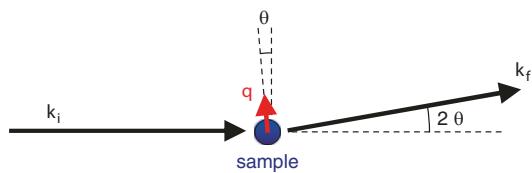
Small-Angle Scattering of Nanostructures and Nanomaterials, Table 1 List of symbols used throughout this entry

θ	Bragg angle of diffraction
\vec{k}_i, \vec{k}_f	Wavevector of incoming, final (scattered) beams
λ	Wavelength of beam
\vec{q}, q	Scattering vector, its magnitude
ζ	Angle of q in detector plane
d	Spacing between Bragg planes
D	Characteristic length scale within sample
\vec{r}, r	(Real space) position vector in sample or within nano-object, its magnitude
b, b_x, b_n, b_m	Scattering length, that of X-rays, that of neutrons from nuclei, that of neutrons from atomic moments
ρ, ρ_0	Scattering length density, that of solvent or matrix
F	Form factor of single nano-object
I	Scattered intensity
M, N	Magnetic, nuclear parts of (neutron) form factor of single nano-object
P	Form factor $P = F ^2$
S	Structure factor
C	Component of magnetic moment perpendicular to \vec{k}_i
ϕ	Angle of magnetic moment in detector plane
Z	Component of magnetic moment parallel to \vec{k}_i
\vec{A}	Halpern-Johnson vector

of crystalline diffraction peaks. Nevertheless, Bragg's law encapsulates a fundamental relationship

$$\lambda = 2d \sin \theta = \frac{4\pi}{q} \sin \theta \quad (1)$$

showing that the angle of diffraction θ varies inversely with the separation d of the diffracting lattice planes. It can be seen that the scattering from objects on nano- to micrometer scales falls predominantly in the small-angle regime $<1^\circ$. A beam of $\lambda \approx 7 \text{ \AA}$ wavelength, for example, would scatter from objects with a d -spacing of $\approx 1000 \text{ \AA}$ into an angle $2\theta \approx 0.4^\circ$ separated from the unscattered beam. Practical constraints on the wavelengths used are imposed by the spectrum of the source available and by the absorption



Small-Angle Scattering of Nanostructures and Nanomaterials, Fig. 1 Schematic depicting the relation for the scattering vector $\vec{q} = \vec{k}_f - \vec{k}_i, 2\theta$ is the angle between the incident beam \vec{k}_i and the scattered beam \vec{k}_f

of the beam by the sample, which increases with wavelength for both X-rays and neutrons. Typically, "soft" X-rays ($\approx 1 \text{ \AA}$, or equivalently, 12 keV) or neutrons ($\approx 5 \text{ \AA}$) from a cold ($\lesssim 40 \text{ K}$) moderator are used in small-angle studies. All scattering measurements are made in reciprocal (Fourier transform) space. The scattering vector \vec{q} , whose magnitude q satisfies (Eq. 1), parameterizes the scattering pattern in reciprocal space and is helpful when comparing patterns between different instruments. The vectorial version of (Eq. 1), $\vec{q} = \vec{k}_f - \vec{k}_i$, is illustrated in Fig. 1, where \vec{k}_i , \vec{k}_f are, respectively, the wavevectors of the incident and final (scattered) beams.

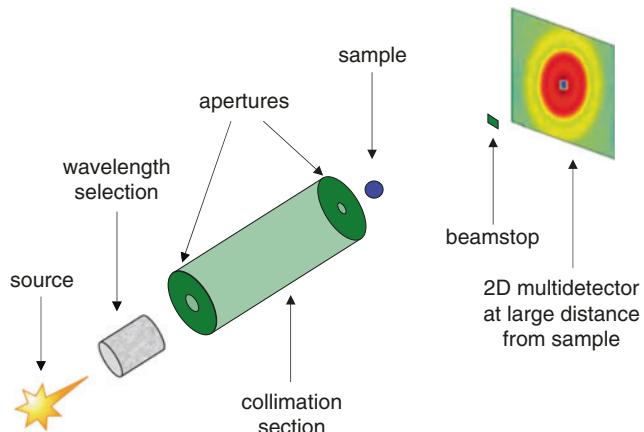
Experimental Method

Small-Angle Scattering

The fundamental task of a small-angle instrument is to separate the intense, unscattered direct beam from the weaker scattering at small angles 2θ . A common experimental setup, depicted in Fig. 2, resembles the classic pinhole camera. Following the source, the wavelength λ of the incoming beam is selected. Due to the low scattering angles, beam intensity can be maintained using a large wavelength spread $\Delta\lambda/\lambda$ without significant detriment to the q -resolution. On small-angle X-ray scattering (SAXS) instruments, a monochromator crystal such as Si (111) is used if a wavelength spread narrower than the source distribution is required. On small-angle neutron scattering (SANS) instruments at pulsed neutron sources, a neutron's wavelength may be calculated by timing its journey to the detector

Small-Angle Scattering of Nanostructures and Nanomaterials

Fig. 2 Schematic of a typical small-angle scattering instrument



(known as the *time of flight* technique). At continuous (e.g., reactor) neutron sources, velocity selectors are mainly used. Velocity selectors comprise of channels or slotted disks forming helical pathways about an axle; only neutrons of a particular velocity are able to pass through the rotating helical pathway. The desired neutron wavelength is selected by varying the rotational speed. The opening angle of the helical pathway is typically set to give wavelength spreads of $\Delta\lambda/\lambda \approx 0.06$.

After the wavelength selection, a collimation section that may consist of a pair of apertures or a slit system ensures that the beam has a tight angular spread $<0.1^\circ$ when it is incident on the sample. The precise collimation and resulting angular spread are often tuned to match the desired measurement range in 2θ (or, equivalently, in q) and to optimize beam intensity. A two-dimensional multidetector, comprising an array of pixels in a plane normal to the unscattered beam, is placed at a sufficient distance behind the sample so as to resolve the small-angle scattering in the desired q range. For example, the D11 SANS instrument at the Institut Laue-Langevin in Grenoble has a multidetector of 128×128 pixels, each pixel $7.5 \times 7.5 \text{ mm}^2$, and features maximum collimation and sample-to-detector distances of 40.5 m and 39 m, respectively, a configuration that permits scattering vectors as low as $q \approx 0.0003 \text{ \AA}^{-1}$ to be measured. The 2D area detectors used in SANS or SAXS are designed to have high detection efficiencies ($>80\%$), and may need to be shielded from intense fluxes, namely, the

unscattered beam, that may saturate and can even damage the detector. A beamstop may be placed in front of the detector for this purpose.

Ultra-Small-Angle Scattering

For some systems one would like to extend the range of observable structures to length scales larger than those which may be probed using the habitual pinhole setup (Fig. 2) and measure scattering at “ultra” small angles. To achieve this, an experimental geometry developed by Bonse and Hart [7], which is now commonplace, utilizes a pre-sample monochromator comprising of a single crystal out of which a channel has been precisely cut. The beam passes in through one end of the channel at an incident angle θ , and appears at the other end (also at an angle θ) after having bounced multiple times from Bragg reflections off both sides of the channel. The resulting beam has an extremely small angular divergence $\lesssim 10^{-4}$ degrees. A similar channel-cut crystal placed behind the sample analyzes the angles of the scattered beam. With this setup, measurements are possible very close to the unscattered beam, down to scattering vectors $q \lesssim 10^{-5} \text{ \AA}^{-1}$, in other words, up to length scales in the tens of micrometer range. Unlike the pinhole SANS setup where a single measurement with the 2D detector captures an entire surface of scattering vectors \vec{q} , the Bonse-Hart geometry can only probe one point in \vec{q} at a time and the measurement times required with this technique are accordingly much longer.

General Theory

In this and in the following subsection (section “[Spherically Symmetric Forms](#)”) the concepts of *scattering length density*, *contrast*, *structure factor*, *form factor*, *polydispersity*, and *Guinier* and *Porod regimes* are introduced. These are also explained in several introductory textbooks [1, 3, 6]. The q -dependence of the measured scattered intensity is described by the “*differential cross section*,” denoted simply as $I(\vec{q})$. For now the scattering is presumed to be elastic ($|\vec{k}_i| = |\vec{k}_f|$ i.e., there is no energy transfer between beam and sample) and coherent (produces interference effects). Detailed treatments of inelastic and incoherent scattering may be found in the textbooks on diffraction methods [8].

Born Approximation and Scattering Length

One presumption of particular pertinence to small-angle scattering experiments is the *Born approximation*, in which the scattering is regarded to be weak (such that the scattering potential may be treated analytically as a perturbation). Within the Born approximation, the photon or neutron scatters no more than once during its passage through the sample. Then the measured scattered intensity is

$$I(\vec{q}) \sim \left| \sum_j b_j e^{i\vec{q} \cdot \vec{R}_j} \right|^2 \quad (2)$$

where \vec{R}_j denotes the position of the j th scatterer in the sample, and b_j is the *scattering length*, epitomizing the capability of the j th object to scatter the beam. Neutrons scatter from nuclei and from magnetic moments [8], while X-rays scatter from electrons (Thomson scattering) [1, 3]: the scattering length of one electron is 2.8×10^{-5} Å. The differences between neutron and X-ray scattering will be discussed in section “[X-rays or Neutrons](#)”. Here and throughout this work, leading prefactors are omitted for simplicity. These prefactors would, for example, ensure that the scattered intensity scales with the illuminated sample volume. In practice $I(\vec{q})$ is scaled onto absolute units through the measurement of

either a standard sample or the incident beam flux, together with corrections for the attenuation of the beam by the sample and instrument-dependent factors such as electronic noise and detector efficiency [9]. With $I(\vec{q})$ converted to absolute units, intensities obtained on different small-angle instruments may be directly compared, and sample properties, for example, the stoichiometry of that part of the sample giving rise to scattering, may be estimated quantitatively.

Most model functions for small-angle scattering data presuppose that the Born approximation holds. In practice, for strong scatterers, care should be taken to verify that this applies before fitting models to $I(\vec{q})$. If multiple scattering is a problem, the probability of scattering may be diminished by making the sample thinner or, where possible, diluting it or decreasing the *contrast*. For grazing incidence experiment geometries such as the grazing-incidence SAXS (GISAXS) technique that are used to resolve in-plane structures on films [1], the Born approximation is sufficient for very small or thin scatterers, otherwise quantitative analysis of the GISAXS profile requires what is known as the “distorted-wave” Born approximation, where the scattering potential is divided into two parts such that an exact solution to the scattering problem may be obtained [10]. The grazing incidence geometry is not expounded here; further details may be found in Refs. [1] and [10].

Scattering Length Density and Contrast

Theoretical small-angle scattering profiles can be calculated directly from models comprised of individual atomic scatterers, but for fitting purposes the computational effort required is too great at present. As the objects leading to small-angle scattering are generally on larger scales, a more continuous, “coarse-grained” description of the scattering potential landscape would seem appropriate. A continuous function $\rho(\vec{r})$, known as the *scattering length density*, is accordingly defined [6]:

$$\rho(\vec{r}) = \sum_j b_j \delta(\vec{r} - \vec{R}_j)$$

and the measured scattered intensity becomes

$$I(\vec{q}) \sim \left| \int e^{i\vec{q} \cdot \vec{r}} \rho(\vec{r}) d\vec{r} \right|^2 \quad (3)$$

$$\sim \left| \int e^{i\vec{q} \cdot \vec{r}} \rho_0 d\vec{r} + \int e^{i\vec{q} \cdot \vec{r}} (\rho(\vec{r}) - \rho_0) d\vec{r} \right|^2 \quad (4)$$

The last equation serves to illustrate the concept of *contrast*. Here it is useful to consider the scattering from nanoparticles in solution, or equally, nanoscale heterogeneities in a solid matrix. ρ_0 would be the scattering length density of the matrix or solvent, while $\rho(\vec{r})$ would be that of the nanoparticles or heterogeneities. The first term in (Eq. 4) involves just the matrix or solvent and is nonzero only at $q = 0$ [the Fourier transform of a constant is $\delta(0)$], so no small-angle scattering originates from the matrix or solvent. In the second term in (Eq. 4), the particles or heterogeneities give forth to small-angle intensity proportional to the squared difference $(\rho(\vec{r}) - \rho_0)^2$. This is referred to as the “contrast factor.” Systematic alterations of the solvent ρ_0 may thus be used to determine the scattering length density of a solute ρ_1 , since the intensity will go to zero at the match point where $\rho_1 = \rho_0$. This is called “contrast matching.” In multicomponent systems, separate contrast-variation analyses for different q -regimes or for different features in the scattering profile can also be used with great effect to reveal the sizes and shapes of the underlying components. Contrast methods were recently demonstrated, for example, in a prominent study of the nanoscale phases within cement [11].

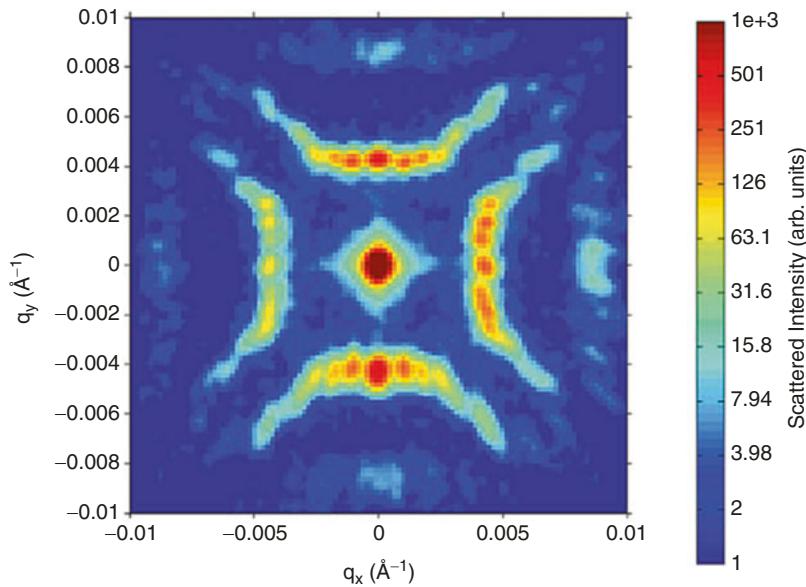
Structure Factor and Form Factor

The notions of *structure factor* and *form factor* are familiar concepts in crystallographic work [8]. Considering N identical scattering nano-objects, each centered at a position \vec{s}_j , the scattered intensity becomes

$$I(\vec{q}) \sim \left| \sum_j^N e^{i\vec{q} \cdot \vec{s}_j} \right|^2 \times \left| \int_{\text{nano-object}} e^{i\vec{q} \cdot \vec{r}} \rho(\vec{r}) d\vec{r} \right|^2 \\ = S(\vec{q}) \times |F(\vec{q})|^2$$

The left side of the multiplication sign defines the *structure factor* $S(\vec{q})$. The integral on the right side ranges over an isolated nano-object, and this term defines the single particle *form factor* $F(\vec{q}) = \int e^{i\vec{q} \cdot \vec{r}} \rho(\vec{r}) d\vec{r}$. It is also commonplace to refer to the function $P(\vec{q}) = |F(\vec{q})|^2$ as the form factor too. In the limit where the nano-objects arrange into a crystal with positions \vec{s}_j forming a perfect lattice, the structure factor becomes a series of δ -functions $S(\vec{q}) = \sum_i \delta(\vec{q} - \vec{G}_i)$, where \vec{G}_i are the reciprocal lattice vectors of the crystal [8]. Crystalline Bragg-diffraction peaks appear in the scattering profile, as illustrated in Fig. 3, with intensities modulated by the form factor. To measure each \vec{G}_i peak, the sample must be rotated with respect to the neutron beam in order to satisfy the condition $\vec{q} = \vec{G}_i$. For situations in the opposite limit where nano-objects are well separated, with no interactions between nano-objects, $S(\vec{q}) = 1$ for all \vec{q} and the small-angle intensity depends only upon $P(\vec{q})$, that is to say, upon the form of an isolated particle. Such a situation is called the “dilute” limit because, with small-angle scattering from solutions, it may be attained by sufficient dilution of the solute [4, 9]. The near-dilute situation is also often realized, where the interparticle interactions are finite but small, such that $S(\vec{q})$ is essentially uniform in the q -range of interest. In situations where the structure factor and the form factor are comparable, for example where the volume fraction of nano-objects is high but no long-range order emerges, the $I(\vec{q})$ must be fitted with models for both $S(\vec{q})$ and $P(\vec{q})$. Details of the many model functions available may be found in the literature [1–6, 13–15].

There are also several numerical methods for modeling $P(\vec{q})$ or $S(\vec{q})$. Numerical methods often start from a radial pair distribution function, which can be obtained by an indirect Fourier transform of $I(q)$ [3]. One approach, commonly used to reconstruct the shape of biological macromolecules, imitates the scattering length density profile $\rho(\vec{r})$ of the macromolecule by several 100 close-packed “beads,” each of constant ρ . Genetic algorithms and simulated annealing



Small-Angle Scattering of Nanostructures and Nanomaterials, Fig. 3 Example of a crystal diffraction pattern on a small-angle neutron detector, here from a flux line lattice [12] in a single crystal of superconducting niobium. To obtain this picture, several detector images were summed as the sample was rotated over a range $\pm 1^\circ$ of rocking angles about the vertical and horizontal axes.

techniques are able to refine the virtual bead assemble to fit the experimental data [4, 16]. Reverse Monte Carlo methods have also been used to recreate the structure factor $S(\vec{q})$ in situations where quasi-long-range order is observed [17].

Spherically Symmetric Forms

Most texts on small-angle scattering begin by considering a spherically centrosymmetric nano-object [3, 6, 13], in other words, a nano-object whose scattering length density $\rho(\vec{r})$ depends only on $r = |\vec{r}|$. Here this scenario is also pursued in order to illustrate concepts such as contrast and polydispersity and furthermore, to highlight the different *regimes* in the scattering profile. To obtain the form factor for a spherical object, it is useful to select a spherical polar coordinate system such that the pole lies along \vec{q} ; the angular components of the integral within $F(\vec{q})$ may then be evaluated, yielding

Here 28 distinct first order Bragg spots are discernible, due to four scalene triangular and two square flux line lattices coexisting in different regions of the sample. The peaks on the left side of the picture appear weaker due to detector efficiency. The unscattered beam, not covered up by a beamstop in this experiment, appears in the center at $q = 0$

$$F(q) = 4\pi \int \rho(r)r^2 \frac{\sin(qr)}{qr} dr$$

The form factor is here a function only of the magnitude $q = |\vec{q}|$ of the scattering vector. Even when the nano-object is not spherical, but a large number of randomly oriented nano-objects are contained in the sample, these orientations may be averaged over and the result becomes again a function of $q = |\vec{q}|$. It is usual to average the scattering measured on the 2D detector (cf. Fig. 2) over the detector's azimuthal direction, so reducing the scattering profile to one dimension, i.e., an I versus q dataset. In section “Magnetic Neutron Scattering” an example is given of where the scattering shows an azimuthal dependence arising for a reason other than crystalline order (cf. Fig. 3). The azimuthal information can impart vital clues as to the nature of the underlying nanostructures.

Core-Shell Form Factor

Core-shell nano-objects are frequently encountered in small-angle scattering studies, e.g., aerosol droplets, polymer micelles [2], inorganic nanoparticles [18]. In addition, core-shell models may provide a useful quantitative starting point even when nano-objects depart from being spherically symmetric. The spherical core-shell model form factor is

$$\begin{aligned} F(q) &= 4\pi \int_0^{r_c} \rho_c r^2 \frac{\sin(qr)}{qr} dr + 4\pi \int_{r_c}^{r_s} \rho_s r^2 \frac{\sin(qr)}{qr} dr \\ &= 4\pi(\rho_c - \rho_s)r_c^3 \frac{j_1(qr_c)}{qr_c} + 4\pi\rho_s r_s^3 \frac{j_1(qr_s)}{qr_s} \end{aligned} \quad (5)$$

where the sphere has a uniform core of radius r_c and scattering length density ρ_c , encapsulated by a uniform shell of scattering length density ρ_s extending to radius $r_s > r_c$, and $j_1(x) = \frac{\sin x}{x^2} - \frac{\cos x}{x}$ is a spherical Bessel function. For brevity, ρ_c and ρ_s are measured relative to the scattering length density ρ_0 of the surrounding solution or matrix; equivalently, ρ_0 may be considered to be zero. Equation 5 may be simplified by introducing volume terms such as $V_c = \frac{4}{3}\pi r_c^3$. Figure 4b exemplary $P(q) = |F(q)|^2$ curves for this core-shell model are plotted.

Polydispersity

In reality, most scattering systems show a distribution in their sizes and shapes; this is known as *polydispersity*. Form factors can be straightforwardly modified to average over a distribution in size or shape, and several functions defining distributions for the polydispersity are available [19]. The effects of polydispersity on the small-angle scattering profile are now examined briefly. In Fig. 4c the core-shell model is modified to include a distribution in core radii with standard deviation 20 % of the mean core radius. By comparing this figure with Fig. 4b, the adverse effects of polydispersity – a smearing in q of the scattering profile – can readily be seen. A smearing of the scattering profile also results from a finite instrument resolution; nonetheless it is clear that sizeable benefits are gained if samples for small-angle

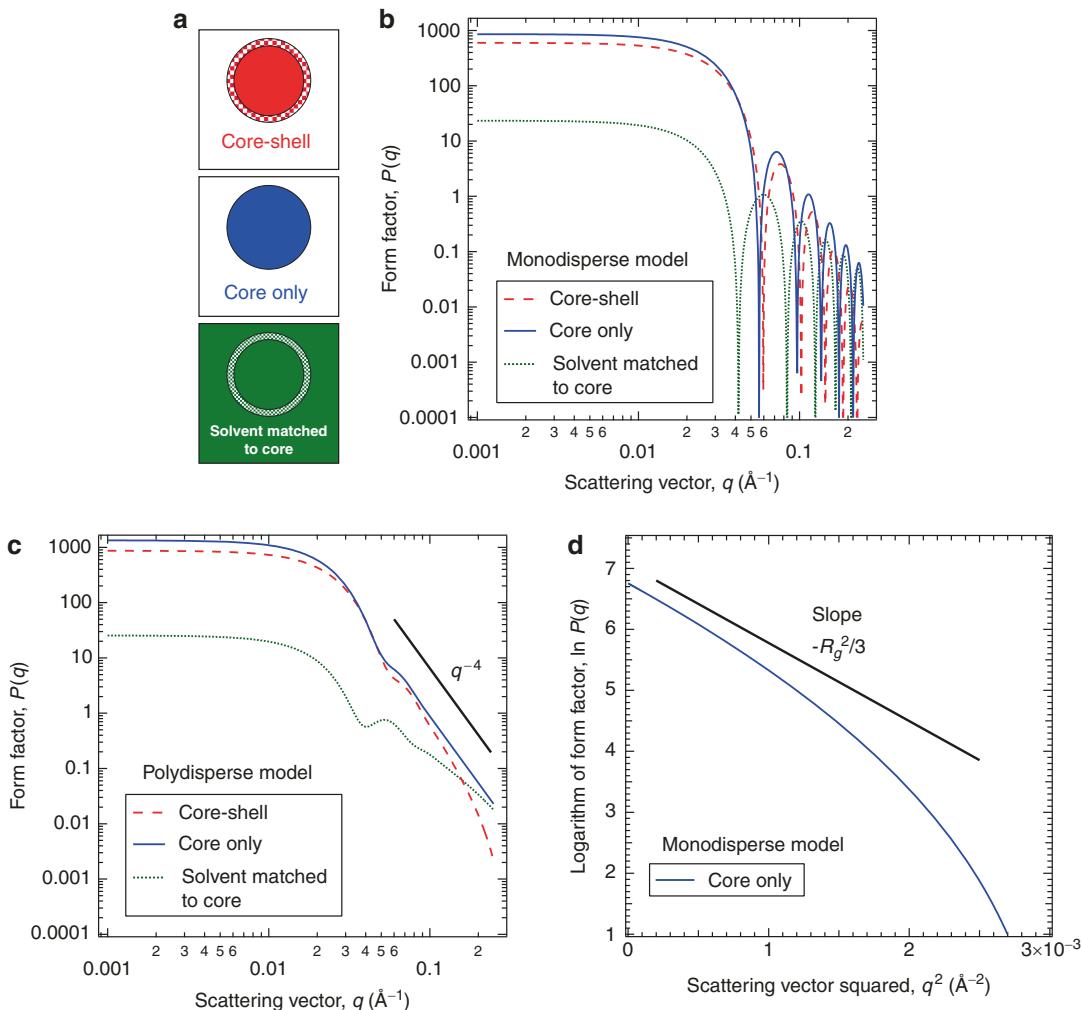
scattering studies are as monodisperse as possible. Figure 4 furthermore illustrates how contrast variation may be exploited, where feasible: by matching the solvent ρ_0 to the core ρ_c (dotted green line in Fig. 4), the interference bumps in the profile are more discernible, notably in polydisperse situations.

Regimes in the Scattering Profile

The appearance of distinctive q regimes in the scattering profile may also be seen in Fig. 4. The solid sphere model (continuous blue lines in Fig. 4) features a characteristic length scale D , namely, the sphere radius 80 Å. The scattering features a high- q regime for $qD \gg 1$, where the scattering profile appears to decay algebraically with q , and a regime at low q where the profile falls off more slowly in q . These regimes emerge independently of the absolute intensity and of any model, and invariably provide useful clues as to the size and homogeneity of the nano-object. In the high- q regime $qD \gg 1$, known as the *Porod regime*, the exponent α of the power-law decay $I(q) \sim q^{-\alpha}$ reveals the “fractal dimension” of the scattering objects [14]. The $\alpha = 4$ observed in Fig. 4c, for example, indicates smooth surfaces, as expected for the model of uniform spheres with boundaries sharp in $\rho(r)$.

In the *Guinier regime* as $q \rightarrow 0$, the scattering follows a Gaussian dependence $\sim \exp(-\frac{1}{3}q^2 R_g^2)$ where Rg is the “radius of gyration” of the nano-object [6], and is analogous to the radius of gyration in mechanics. For uniform solid spheres of radius r_c , $R_g^2 = \frac{3}{5}r_c^2$. To reveal this behavior at low q , a “Guinier plot” is useful, where the logarithm of the scattering is plotted versus q^2 ; the plot will be linear with slope $-\frac{1}{3}R_g^2$ at sufficiently low q . A Guinier plot for the sphere model is shown in Fig. 4d, together with the expected low- q slope (black line in Fig. 4d). Comparing the two, it may be seen that the Guinier approximation holds reasonably well when $q^2 \lesssim 0.0008 \text{ \AA}^{-2}$, that is, for $qRg \lesssim 1.3$. The upper limit for the Guinier regime is found to depend somewhat on the scattering model [13].

Almost every small-angle scattering article incorporates an analysis of the low- or high- q



Small-Angle Scattering of Nanostructures and Nanomaterials, Fig. 4 The spherical core-shell form factor $P(q)$. (a) Schematic showing the various parameters for the model curves plotted in (b–d): in (b) the particles are monodisperse; in (c) the particles have polydisperse core sizes with the standard deviation of the polydispersity distribution set at 20 % of the mean core radius r_c . The dashed red and dotted green lines indicate

$P(q)$ with $r_c = 70 \text{\AA}$ and a shell 10 \AA thick; the dashed red line shows $P(q)$ when the core scattering length density ρ_c is twice that of the shell ρ_s ; the dotted green line shows $P(q)$ when the solvent ρ_0 matches ρ_c . The continuous blue line shows $P(q)$ for a core only model with $r_c = 80 \text{\AA}$; in (d), this line is redrawn on $\ln P(q)$ versus q^2 axes. Black lines are guides to the eye

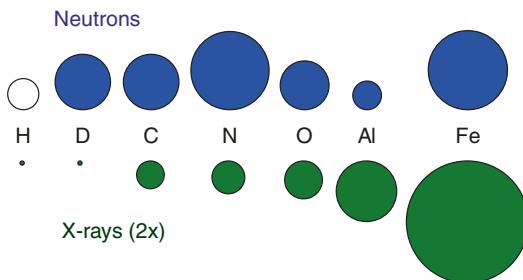
regimes. Extensive examples of Guinier plots and associated methods (e.g., the Zimm and Kratky plots) that quantify scattering regimes without resorting to detailed model fitting may also be found in the literature [3]. There have also been efforts to encapsulate both low- q and high- q regimes in unified models. A common example is the Beaucage model, applicable to situations where a nanostructure exhibits a hierarchy of scales, each

discernible scale contributing a pair of Guinier and Porod regions to the scattering profile [15].

X-rays or Neutrons?

Scattering Length

The basic principles of small-angle scattering, as overviewed in previous sections, have the same



Small-Angle Scattering of Nanostructures and Nanomaterials, Fig. 5 Schematic indicating the propensity for scattering of X-rays or neutrons [20] by selected atoms. The radii of the circles are drawn proportional to the scattering length. For X-rays, the scale has been 2× magnified for clarity. The scattering length of hydrogen for neutrons is negative, in contrast to the other atoms depicted

form for both X-rays and neutrons. However, there are fundamental distinctions between the techniques arising from how the X-rays or neutrons interact with matter. One such distinction is patent in the atomic scattering length b [cf. (Eq. 2)]. For neutrons, b_n depends on the structure of the nucleus and is isotope dependent. For X-rays, b_x depends on the electronic structure and scales with the total number of electrons Z in the atom. At the length scales probed by small-angle scattering, nuclei look like point sources of scattering and b_n is q -independent, whereas for X-rays (atomic absorption edges neglected) $b_x(q)$ falls off slowly with q – this is essentially the Guinier regime of the form factor for the atomic electron clouds, with $b_x(0) = Z \times 2.8 \times 10^{-5} \text{ \AA}$. Low-resolution modeling of SAXS profiles can be achieved with b_x considered constant. The scattering lengths for selected atoms are pictured in Fig. 5 for both neutrons [20] and X-rays.

Generation of Contrast

The scattering length density ρ of an elementary volume V (e.g., a molecular volume) within the sample is calculated by summing the scattering lengths of atoms within V , that is, $\rho = \sum_j b_j / V$. Here, a uniform electron density needs to be assumed for X-rays. The scattering length densities for neutrons of selected substances is listed in Table 2. Comparing with Fig. 5, it is clear that there is an enormous distinction between

Small-Angle Scattering of Nanostructures and Nanomaterials, Table 2 The scattering properties, for neutrons, of selected substances

Substance	Neutron attenuation length (cm)	Neutron scattering length density ($\times 10^{-6} \text{ \AA}^{-2}$)
H ₂ O	0.18	-0.56
D ₂ O	1.55	6.33
C	1.59	7.53
B	0.0036	6.91
¹⁰ B	0.00067	-0.14
¹¹ B	1.35	8.51
Al	7.74	2.08
Fe	0.629	8.02
Cd	0.0031	2.27

For the calculation of these values, room temperature densities are used, and the elements with no isotope specified are deemed to comprise of their isotopes in naturally abundant ratios

hydrogen and deuterium. The former has $b_n = -3.74 \times 10^{-5} \text{ \AA}$, the latter $6.67 \times 10^{-5} \text{ \AA}$ [20], and this is reflected in ρ as shown for H₂O and for D₂O in Table 2. Great contrast may be achieved by using hydrogenated solutes in deuterated solvents. Polystyrene in toluene, for example, presents a contrast of $0.48 \times 10^{-6} \text{ \AA}^{-2}$ when both are hydrogenated, whereas the contrast of hydrogenated polystyrene in deuterated toluene is $4.2 \times 10^{-6} \text{ \AA}^{-2}$. In contrast-variation studies, tuning the solvent scattering length density ρ_0 may be accomplished by using a binary mixture whose components have widely different scattering length densities; these are mixed in a linear ratio to yield the desired ρ_0 . For X-rays, the two components would have different electron densities; for neutrons, they may be deuterated and hydrogenated versions of the same solvent. Isotope substitution may also be used to create SANS contrast *inside* a particle by specific isotope labeling of a particular part of interest, which can be used, for example, to determine how this part fits in with the rest of the particle. In SAXS, a novel way to induce contrast is to measure scattering profiles at different photon energies in the vicinity of an atomic absorption edge of one of the elements within the sample: for this element, the form factor b_x varies across the absorption edge, while the rest of the scattering potential landscape is essentially

invariant in the small range of photon energies measured. This contrast-variation technique is known as anomalous SAXS (ASAXS) and may be exploited for the large number of elements having an absorption edge in the range 5–25 keV [1, 9, 21].

Neutron Absorption

Another aspect of neutron scattering that is heavily dependent on the isotope is absorption. A few elements, for example boron and cadmium (c.f. Table 2), are outstanding neutron absorbers in their naturally abundant forms. Not surprisingly, boron-containing substances and cadmium are common shielding materials. For scattering explorations of these and other such materials, samples need to be prepared using elements devoid of the absorbing isotopes; boron-11, for example, is seen from Table 2 to have an acceptable attenuation length. Separating elements into their different isotopes, however, becomes increasingly difficult and expensive for heavier atoms. It is also apparent from Table 2 that some materials, such as aluminum, are rather poor neutron absorbers. This is providential for the construction of bulky or demanding sample environments such as cryostats or pressure cells.

General Considerations

The majority of elements have neutron attenuation lengths on the order of a centimeter, reflecting the weak interaction of neutrons with matter. In general, neutron scattering explores the bulk of samples and, for similar reasons, effectively no radiation damage to the sample occurs with the neutron fluxes currently available (up to $\approx 10^8$ neutrons $\text{cm}^{-2} \text{s}^{-1}$) on SANS instruments. For SAXS experiments at synchrotron sources, the high brilliance of the X-ray beam ($\approx 10^{16}$ photons $\text{cm}^{-2} \text{s}^{-1}$) can damage soft matter and biological samples within milliseconds [9]. This may affect the validity of the scattering profile measurement. For SAXS the sample thickness may also need to be chosen to obtain the optimum scattered intensity, since at the photon energies (≈ 10 keV) used in SAXS, the X-ray attenuation length is on the order of millimeters for the lighter elements [9], which make up most soft matter systems. The many orders of magnitude more intense fluxes

available on SAXS beamlines are, of course, invaluable for high-resolution, low uncertainty measurements of scattering profiles. The associated fast data acquisition times also facilitate the characterization of time-dependent effects. Likewise, relatively small sample quantities are typically required (<50 μL for SAXS) compared to SANS (≈ 1 mL).

Neutron Incoherent and Spin Scattering

In previous sections, it has been shown that variations in scattering length density on the nanoscale give rise to coherent small-angle scattering. However, if the scattering length b varies in a disordered fashion for different atoms of the same element, then for neutrons, since the nuclei are effectively point sources, this gives rise to incoherent (q -independent) scattering [8]. This adds to the background of the coherent scattering profile and in a few cases may be significant enough to limit the measurement even when longer count times are used. A variation in b for the same element can arise due to different isotopes, present in sufficient abundance, having widely different scattering lengths. Another reason for a variation in b occurs when the nuclei have nonzero spin and the scattering length depends on the spin state. Neutrons are sensitive to such spin disorder because the neutron itself has spin $\frac{1}{2}$. A notorious example is hydrogen: this has an incoherent scattering length of 25.3×10^{-5} Å, compared to 4.1×10^{-5} Å for deuterium. Deuterated solvents should therefore be used where possible.

The spin of the neutron is rather advantageous for the exploration of magnetic structures. Small-angle scattering can arise from nano- to micrometer-scale variations in the magnetic field, as well as correlations over similar length scales between atomic moments. This is discussed in the following section.

Magnetic Neutron Scattering

The spin- $\frac{1}{2}$ nature of the neutron and its associated magnetic moment make neutron scattering techniques a natural probe of magnetic systems. Indeed, magnetic structures are considerably better

understood following the dawn of high-flux neutron sources. The origin of an observed scattering signal – whether it is nonmagnetic (i.e., from nuclei) or magnetic in origin – can be determined by measurements at either side of the appropriate thermodynamic phase transition. In other cases, for example if magnetic and structural order appear simultaneously, the origin of the signal may be determined by analyzing the neutron polarization and/or the azimuthal dependence of the scattering on the 2D detector. Furthermore, neutron polarization analysis can convey additional information as to the orientation of the magnetic moments behind the scattering signal. The small-angle applications of neutron polarization analysis are introduced in the second part of this section (section “[Neutron Polarization Analysis](#)”).

Unpolarized Neutrons

The interaction potential of the neutron with a magnetic field \vec{B} is $-\vec{\mu} \cdot \vec{B}$, where $\vec{\mu} = \gamma\mu_N \vec{\sigma}$ is the magnetic moment operator for the neutron, γ is the neutron gyromagnetic ratio, and μ_N is the nuclear magneton [8]. $\vec{\sigma}$ is the Pauli spin operator describing the neutron spin state; its average value characterizes the polarization of the neutron beam. All the general principles explained in the previous sections can be used to describe the small-angle scattering from magnetic structures, whether the neutrons are polarized or not. Noting, for example, that the scattering interaction potential $-\vec{\mu} \cdot \vec{B}$ is already continuous, (Eq. 3) for the unpolarized elastic coherent scattering becomes, with a collinear field B in the sample

$$I(\vec{q}) \sim \int e^{i\vec{q} \cdot \vec{r}} B(\vec{r}) d\vec{r}$$

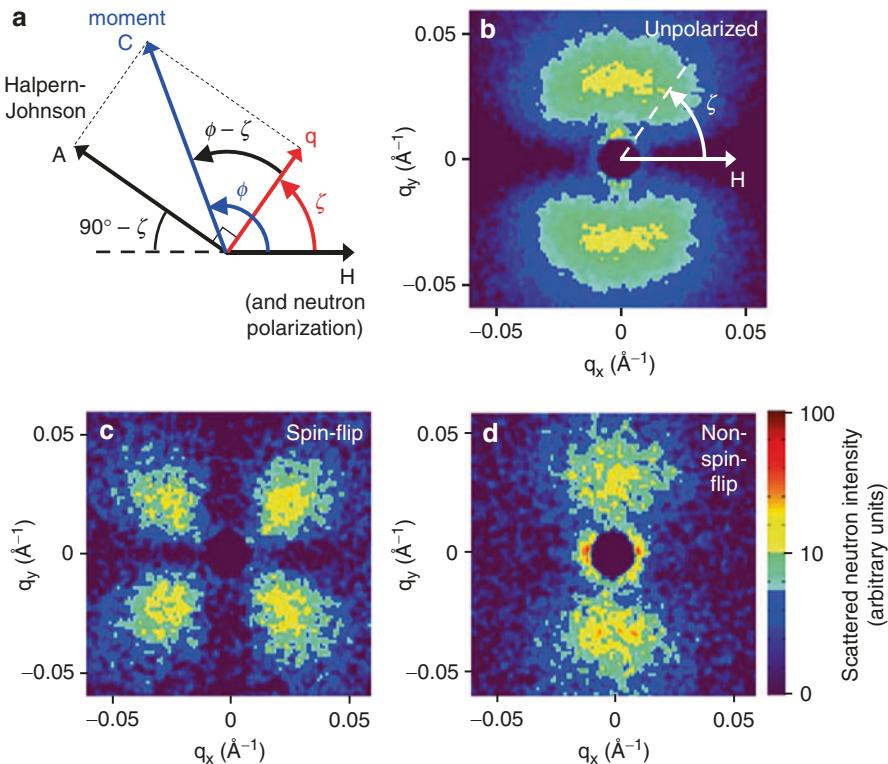
This prescription is very convenient for flux line lattices that form in Type-II superconductors [12], where crystalline diffraction peaks are frequently observed, as illustrated in Fig. 3. The long-range order and nanometer d -spacings of magnetic skyrmion lattices also result in SANS diffraction peaks [22].

The magnetic scattered intensity may also be estimated in terms of magnetic scattering lengths of individual atoms. Like the atomic form factors

in SAXS, the magnetic scattering length $b_m(q)$ is close to constant at low- q [8], with $b_m = m \times 2.7 \times 10^{-5} \text{ \AA}/\mu_B$ where m is the atomic moment in Bohr magnetons μ_B [5, 23]. Under the proviso that the model chosen to fit the scattered intensity is representative, this relation can be utilized to estimate the underlying moment behind a scattering feature. With SANS, studies may be directed at the magnetism of individual atoms, since in ferri- and ferromagnetic systems in particular, the structure factor at low- q affords a measure of the interactions between individual atomic moments separated by large distances. An example is the collection of “random anisotropy” magnets, where SANS arises from random variations in the magnetic anisotropy fields. SANS studies on these materials are the subject of a recent review [5], in which various structure factor models pertinent to magnetic scattering are also detailed.

Magnetic domain walls in ferri- or ferromagnets also give rise to small-angle scattering. Such walls are normally observed to be spatially uncorrelated so that the scattering is dominated by the form factor. Because the domain sizes D are typically micrometers or above, a pinhole SANS instrument is able only to measure the high- q (Porod) regime of this scattering. The full profile may be accessible with an ultra-small-angle instrument [7, 24]. An example of where structure-factor-dominated and form-factor-dominated magnetic scattering features occur in the same sample is demonstrated in the phase-separated perovskite oxides, which exhibit colossal magnetoresistance [25]. In these systems, micrometer-sized ferromagnetic regions materialize in the nonmagnetic matrix below a critical temperature T_c . The Porod regime of their form factor appears as SANS intensity $I(q) \sim q^{-4}$. At the same time, at temperatures in the vicinity of T_c , quasielastic critical fluctuations give (Ornstein–Zernike) scattering, which has a Lorentzian profile $I(q) \sim 1/[q^2 + (1/\xi)^2]$ where ξ is the correlation length.

The dipole nature of the interaction between the neutron and atomic magnetic moments means that only moments \vec{m} which are perpendicular to the scattering vector \vec{q} are effective in scattering [8, 23]. The relevant part of \vec{m} is expressed by the Halpern–Johnson vector



Small-Angle Scattering of Nanostructures and Nanomaterials, Fig. 6 2D SANS detector images from magnetically distinct nanoscale heterogeneities in a ferromagnet. (a) Schematic drawn in the detector plane depicting how detector anisotropy arises from the component of the moment C lying at an angle ϕ in the detector

plane. (b-d) Measurements from an iron-gallium single crystal under saturating field H applied along $\zeta = 0$ in the detector plane such that $\phi = 0$: (b) unpolarized neutrons, (c) polarized neutrons that scatter with spin flip or (d) without spin flip

$$\vec{A} = \hat{\mathbf{q}} \wedge (\vec{m} \wedge \hat{\mathbf{q}}) = \vec{m} - (\vec{m} \cdot \hat{\mathbf{q}}) \hat{\mathbf{q}}$$

where $\hat{\mathbf{q}}$ is the unit vector $\vec{q} / |\vec{q}|$. In terms of \vec{A} , the unpolarized neutron scattering (Eq. 2) becomes

$$I(\vec{q}) \sim \sum_{j,l} \left(b_{nj} b_{nl}^* + b_{mj} b_{ml}^* \vec{A}_j \vec{A}_l \right) e^{i\vec{q} \cdot (\vec{R}_j - \vec{R}_l)}$$

Spin-dependent (incoherent) scattering from nuclei has been excluded here; this would enter in the form of terms similar to those in A [23]. The Halpern-Johnson relationship can lead to an anisotropy on the SANS detector. To describe this anisotropy, it is convenient to use cylindrical polar coordinates with the longitudinal axis along the beam direction \vec{k}_i . The polar plane of this

coordinate system is drawn schematically in Fig. 6a. The scattering vector $\vec{q} \simeq (q, \zeta, 0)$ at small angles. Magnetically distinct heterogeneities within a ferromagnet provide a useful demonstration: the moment of the j th nanoscale heterogeneity, or rather, the contrast of this nanoscale moment with the ferromagnetic matrix, is described by (C_j, ϕ_j, Z_j) in the cylindrical polar coordinates. The unpolarized small-angle scattered intensity is

$$I(\vec{q}) \sim \sum_{j,l} \left(N_j(\vec{q}) N_l^*(\vec{q}) + M_j(\vec{q}) M_l^*(\vec{q}) (Z_j Z_l + C_j C_l \sin(\zeta - \phi_j) \times \sin(\zeta - \phi_l)) e^{i\vec{q} \cdot (\vec{R}_j - \vec{R}_l)} \right) \quad (6)$$

where $N(q)$ and $M(q)$ are, respectively, the nuclear and magnetic components of the form factor of a

single nanoscale heterogeneity. $(\zeta - \phi)$ is the angle between the moment and \vec{q} ; the component perpendicular to the scattering vector \vec{q} is $\sin(\zeta - \phi)$, as shown in Fig. 6a. Magnetic heterogeneities with moments co-aligned along a direction ϕ in the detector plane will give rise to a $\sin^2(\zeta - \phi)$ anisotropy on the detector, as illustrated in Fig. 6b. Correspondingly, any observed sine-squared anisotropy on the detector could indicate that the small-angle scattering is magnetic in origin. Nuclear scattering can also be anisotropic, for example, if elongated particles are co-aligned in a direction perpendicular to the beam. For further confirmation that the observed scattering is magnetic, analysis of the neutron polarization can be performed.

Neutron Polarization Analysis

A multilayer or “supermirror” (e.g., of Fe and Si) is commonly used to polarize the incident cold neutron beam. A similar device can in principle be employed to analyze the spin of the scattered neutrons. For SANS, however, the angular acceptance of these devices is not sufficient to cover the 2θ range of scattering angles and these devices may also confer an undesirable small-angle background. The recent development of ${}^3\text{He}$ cells that resolve these issues has made it possible to routinely perform polarization analysis on small-angle scattering profiles. Helium-3 has a particularly spin-dependent neutron absorption: with 100 % of the nuclei in the ${}^3\text{He}$ cell polarized, one spin state of the neutrons (that parallel to the ${}^3\text{He}$ spin) is negligibly absorbed, while the other is strongly absorbed. Thus one is able to determine whether, during scattering, a neutron flips its spin (“spin-flip” scattering) or not (“non-spin-flip” scattering). In other words, the polarization is analyzed fully in one direction; this is called “longitudinal” polarization analysis. The scattered intensity breaks down into four components, denoted $++$, $+ -$, $- +$, and $--$; the first symbol designates the incident neutron spin, the second the spin of the scattered neutron.

Longitudinal polarization analysis allows an unambiguous distinction to be made between coherent nuclear scattering and spin (incoherent) nuclear scattering, or between magnetic scattering

and nuclear scattering [21, 23]. The latter is aptly demonstrated with the example of magnetically distinct heterogeneities in a ferromagnetic matrix. The contribution from nuclear spins is excluded, as before. The non-spin-flip channel is found to be sensitive to nuclear scattering and to magnetic scattering that has a component of the Halpern–Johnson vector \vec{A} parallel to the neutron polarization. Meanwhile, all the scattering observed in the spin-flip channel is magnetic in origin, arising from the components of the Halpern–Johnson vector \vec{A} perpendicular ($\vec{A} \perp$) to the neutron polarization. The neutron polarization at the sample follows the applied magnetic field \vec{H} , that is, along $\zeta = 0$ in this example. Then, the polarized scattered intensity

$$I^{\pm\mp}(\vec{q}) \sim \sum_{j,l} M_j M_l^* (Z_j Z_l + C_j C_l \sin(\zeta - \phi_j) \sin(\zeta - \phi_l) \cos^2 \zeta) e^{i \vec{q} \cdot (\vec{R}_j - \vec{R}_l)} \quad (7)$$

$$I^{\pm\mp}(\vec{q}) \sim \sum_{j,l} (N_j N_l^* \pm (M_j N_l^* C_j \sin(\zeta - \phi_j) + N_j M_l^* C_l \sin(\zeta - \phi_l)) \sin \zeta + M_j M_l^* C_j C_l \sin(\zeta - \phi_j) \sin(\zeta - \phi_l) \times \sin^2 \zeta) e^{i \vec{q} \cdot (\vec{R}_j - \vec{R}_l)} \quad (8)$$

where the form factors are \vec{q} -dependent functions $N(\vec{q})$ and $M(\vec{q})$. Comparing (Eq. 7) and (Eq. 8) to the unpolarized scattering (Eq. 6), additional $\sin(90^\circ - \zeta)$ and $\cos(90^\circ - \zeta)$ terms appear in the spin-flip and non-spin-flip channels, respectively. The origin of these terms, as illustrated in Fig. 6a, lies in the sensitivity of polarized neutrons to the components of \vec{A} . In the spin-flip channel (Eq. 7), a $\vec{A} \perp \vec{A} \wedge \vec{A} \perp$ term has been omitted that arises from chiral magnetic structures, such as spin helices or skyrmions [26]. The small-angle scattering from magnetic nanoparticles provides an example of where a significant nuclear component $N(\vec{q})$ of the form factor arises in conjunction with a magnetic component $M(\vec{q})$ from the same nano-object [18]. For this particular example of magnetically distinct heterogeneities illustrated in Fig. 6, the nuclear form factor $N(\vec{q})$ is found to be

negligible. With moments co-aligned at magnetic saturation along $\phi = 0$, the anisotropy on the detector follows a $\sin^4 \zeta$ dependence for the non-spin-flip channel and a $\sin^2 \zeta \cos^2 \zeta$ dependence for the spin-flip channel (Fig. 6c, d). It is clear that these anisotropies have provided unambiguous confirmation that the origin of the small-angle scattering is magnetic.

Summary and Outlook

Small-angle scattering techniques have made an enormous impact on condensed matter research, spanning across several disciplines. In this short entry, a comprehensive review of the wealth of publications in this realm has been avoided. Instead, the basic principles that form the quoin of topical small-angle scattering articles have been overviewed. The advantages and practical aspects of small-angle neutron and X-ray techniques have also been compared. For further reading, in-depth reviews focusing on selected disciplines are available, in particular for studies of soft matter [1, 2, 4, 9, 10]. These include discussions of the applications of small-angle scattering for biology [4], polymers [1, 2], disordered solids [1], and SAXS on solutions in general [3].

The examples presented in this entry tend toward “hard” condensed matter research, as despite significant topical results, there are presently few reviews in this area. Exceptions may be found in Ref. [5], where the focus is on magnetic SANS, in particular for nanocrystalline materials and magnetic fluids. In particular, magnetic neutron scattering and neutron polarization analysis have been introduced. Neutron polarization analysis has only just become routinely available on small-angle neutron instruments due to the recent development of ${}^3\text{He}$ apparatus to analyze the spin of the scattered neutron. This enables the separation of spin-incoherent from coherent scattering [1, 21]. Furthermore, through the particular character of the interaction between neutron spin and atomic moments [23], small-angle polarized neutron studies can provide tremendous insights into the magnetic nature of nanostructures [5, 18].

In conjunction with advances in beamline apparatus, significant progress is also resulting from the availability of increasingly powerful sources of X-rays and neutrons, which allow the imaging of nanostructures in unprecedented ways. The advent of the free electron laser – a source providing super-intense, femtosecond pulses of X-rays – allows unrivaled scattering studies in real time on ultrashort timescales, from the functional dynamics of biomolecules to magnetic spin-flip processes [27]. Concurrently, advancement steadily continues with the enhancement of existing techniques [28] and the introduction of new scattering models and numerical methods [4, 5, 15–17].

Acknowledgments DanScatt is acknowledged for financial support and Annabel J. Lingham is thanked for a careful reading of the manuscript.

Cross-References

Note: With the exception of ‘Synchrotron Radiation Techniques’, these cross-references are essentially a list of applications where it is apparent that small-angle scattering techniques have been employed and made a nontrivial contribution to the field. The length of the list reflects that small-angle scattering provides a staple technique across the realm of nanotechnology, and is routinely used for sample characterization.

- Carbon Nanotubes
- Chitosan Nanoparticles
- Dermal and Transdermal Delivery
- Electric Field-Directed Assembly of Bioderivatized Nanoparticles
- Fullerenes for Drug Delivery
- Gas-Phase Nanoparticle Formation
- Gold Nanorods
- Hollow Gold Nanospheres
- Light-Element Nanotubes and Related Structures
- Liposomes
- Macromolecular Crystallization Using Nano-volumes

- ▶ Magnetic-Field-Based Self-assembly
- ▶ Nano-Concrete
- ▶ Nanomaterials for Electrical Energy Storage Devices
- ▶ Nanomaterials for Excitonic Solar Cells
- ▶ Nanoparticles
- ▶ Nanoscale Water Phase Diagram
- ▶ Nanostructured Thermoelectric Materials
- ▶ Nanostructures Based on Porous Silicon
- ▶ Nanostructures for Coloration (Organisms other than Animals)
- ▶ Nanostructures for Photonics
- ▶ Optical Properties of Metal Nanoparticles
- ▶ Prenucleation Clusters
- ▶ Selected Synchrotron Radiation Techniques
- ▶ Self-assembly
- ▶ Self-assembly of Nanostructures
- ▶ Spider Silk
- ▶ Structural Color in Animals
- ▶ Wetting Transitions

References

1. Naudon, A., Schmidt, P.W., Stuhrmann, H.B.: In: Brumberger, H. (ed.) *Modern Aspects of Small-angle Scattering*, pp. 1–56, 181–220, and 221–254. Kluwer, Dordrecht (1995)
2. Hammouda, B.: SANS from polymers – review of the recent literature. *J. Macromol. Sci. Part C: Polym. Rev.* **50**, 14–39 (2010); Melnichenko, Y.B., Wignall, G.D.: Small-angle neutron scattering in materials science: recent practical applications. *J. Appl. Phys.* **102**, 021101 (2007)
3. Glatter, O., Kratky, O. (eds.): *Small Angle X-ray Scattering*. Academic, London (1982); Feigin, L.A., Svergun, D.I.: *Structure Analysis by Small-Angle X-ray and Neutron Scattering*. Plenum, New York (1987)
4. Putnam, C.D., Hammel, M., Hura, G.L., Tainer, J.A.: X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q. Rev. Biophys.* **40**, 191–285 (2007); Kasai, N., Kakudo, M.: *X-ray Diffraction by Macromolecules*. Springer, Berlin (2005)
5. Michels, A., Weissmüller, J.: Magnetic-field-dependent small-angle neutron scattering on random anisotropy ferromagnets. *Rep. Prog. Phys.* **71**, 066501 (2008); Avdeev, M.V., Aksenov, V.L.: Small-angle neutron scattering in structure research of magnetic fluids. *Phys. – Uspekhi* **53**, 971–993 (2010)
6. Guinier, A., Fournet, G.: *Small-Angle Scattering of X-rays*. Wiley, New York (1955)
7. Bonse, U., Hart, M.: Tailless X-ray single-crystal reflection curves obtained by multiple reflection. *Appl. Phys. Lett.* **7**, 238–240 (1965)
8. Squires, G.L.: *Introduction to the Theory of Thermal Neutron Scattering*. Cambridge University Press, Cambridge (1978); Lovesey, S.W.: *Theory of Thermal Neutron Scattering from Condensed Matter*, vol. 1. Oxford University Press, Oxford (1984); Balcar, E., Lovesey, S.W.: *Theory of Magnetic Neutron and Photon Scattering*. Oxford University Press, Oxford (1989)
9. Narayanan, T., Grillo, I.: In: Borsali, R., Pecora, R. (eds.) *Soft Matter Characterization*, pp. 725–782 and 900–952. Springer, Berlin (2008)
10. Tolan, M.: X-ray Scattering from Soft-matter Thin Films. Springer, Berlin (1999); Okuda, H., Kato, M., Kuno, K., Ochiai, S., Usami, N., Nakajima, K., Sakata, O.: A grazing incidence small-angle x-ray scattering analysis on capped Ge nanodots in layer structures. *J. Phys.: Condens. Matt.* **22**, 474003 (2010)
11. Allen, A.J., Thomas, J.J., Jennings, H.M.: Composition and density of nanoscale calcium-silicate-hydrate in cement. *Nat. Mater.* **6**, 311–316 (2007)
12. Huxley, A.: In: Huebener, P., Schopohl, N., Volovik, G.E. (eds.) *Vortices in Unconventional Superconductors and Superfluids*, pp. 301–339. Springer, Berlin (2002)
13. Glatter, O., May, R.: Small-angle techniques. In: Prince, E. (ed.) *International Tables of Crystallography*, vol. C, pp. 89–112. Kluwer, Dordrecht (2004)
14. Teixeira, J.: Small-angle scattering by fractal systems. *J. Appl. Crystallogr.* **21**, 781–785 (1988); Bale, H.D., Schmidt, P.W.: Small-angle X-ray-scattering investigation of submicroscopic porosity with fractal properties. *Phys. Rev. Lett.* **53**, 596–599 (1984); Ruland, W.: Small-angle scattering of two-phase systems: determination and significance of systematic deviations from Porod's law. *J. Appl. Crystallogr.* **4**, 70–73 (1971)
15. Beaucage, G.: Approximations leading to a unified exponential/power-law approach to small-angle scattering. *J. Appl. Crystallogr.* **28**, 717–728 (1995); Hammouda, B.: Analysis of the Beaucage model. *J. Appl. Crystallogr.* **43**, 1474–1478 (2010)
16. Chacón, P., Fernando Díaz, J., Morán, F., Andreu, J. M.: Reconstruction of protein form with X-ray solution scattering and a genetic algorithm. *J. Mol. Biol.* **299**, 1289–1302 (2000); Svergun, D.I.: Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* **76**, 2879–2886 (1999); Svergun, D.I., Petoukhov, M.V., Koch, M.H.J.: Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.* **80**, 2946–2953 (2001)
17. Laver, M., Forgan, E.M., Abrahamsen, A.B., Bowell, C., Geue, T., Cubitt, R.: Uncovering flux line correlations in superconductors by reverse Monte Carlo refinement of neutron scattering data. *Phys. Rev. Lett.* **100**, 107001 (2008)

18. Krycka, K.L., Booth, R.A., Hogg, C.R., Ijiri, Y., Borchers, J.A., Chen, W.C., Watson, S.M., Laver, M., Gentile, T.R., Dedon, L.R., Harris, S., Rhyne, J. J., Majetich, S.A.: Core-shell magnetic morphology of structurally uniform magnetite nanoparticles. *Phys. Rev. Lett.* **104**, 207203 (2010)
19. Walter, G., Kranold, R., Gerber, T., Baldrian, J., Steinhart, M.: Particle size distribution from small-angle X-ray scattering data. *J. Appl. Crystallogr.* **18**, 205–213 (1985)
20. Sears, V.F.: Neutron scattering lengths and cross-sections. *Neutron News* **3**, 26–37 (1992)
21. Stuhrmann, H.B.: Contrast variation in X-ray and neutron scattering. *J. Appl. Crystallogr.* **40**, s23–s27 (2007)
22. Mühlbauer, S., Binz, B., Jonietz, F., Pfleiderer, C., Rosch, A., Neubauer, A., Georgii, R., Böni, P.: Skyrmion lattice in a chiral magnet. *Science* **323**, 915–919 (2009)
23. Moon, R.M., Riste, T., Koehler, W.C.: Polarization analysis of thermal-neutron scattering. *Phys. Rev. B* **181**, 920–931 (1969); Halpern, O., Johnson, M.H.: On the magnetic scattering of neutrons. *Phys. Rev.* **55**, 898–923 (1939)
24. Wagner, W., Bellmann, D.: Bulk domain sizes determined by complementary scattering methods in polycrystalline Fe. *Physica B* **397**, 27–29 (2007)
25. Wu, J., Lynn, J., Glinka, C.J., Burley, J., Zheng, H., Mithcell, J.F., Leighton, C.: Intergranular giant magnetoresistance in a spontaneously phase separated perovskite oxide. *Phys. Rev. Lett.* **94**, 037201 (2005)
26. Grigoriev, S.V., Chernyshov, D., Dyadkin, V.A., Dmitriev, V., Maleyev, S.V., Moskvin, E.V., Menzel, D., Schoenes, J., Eckerlebe, H.: Crystal handedness and spin helix chirality in $\text{Fe}_1 - x\text{Co}_x\text{Si}$. *Phys. Rev. Lett.* **102**, 037204 (2009); Pappas, C., Lelièvre-Berna, E., Falus, P., Bentley, P.M., Moskvin, E., Grigoriev, S., Fouquet, P., Farago, B.: Chiral paramagnetic skyrmion-like phase in MnSi. *Phys. Rev. Lett.* **102**, 197202 (2009)
27. Treusch, R., Feldhaus, J.: FLASH: new opportunities for (time-resolved) coherent imaging of nanostructures. *New J. Phys.* **12**, 035015 (2010)
28. Hura, G., Menon, A.L., Hammel, M., Rambo, R.P., Poole, F.L., Tsutakawa, S.E., Jenney, F.E., Classen, S., Frankel, K.A., Hopkins, R.C., Yang, S., Scott, J.W., Dillard, B.D., Adams, M.W.W., Tainer, J.A.: Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat. Methods* **6**, 606–614 (2009)

Small-Angle X-Ray Scattering

- Small-Angle Scattering of Nanostructures and Nanomaterials

Smart Carbon Nanotube-Polymer Composites

- Active Carbon Nanotube-Polymer Composites

Smart Drug Delivery Microchips

- Stimuli-Responsive Drug Delivery Microchips

Smart Hydrogels

Alessandro Parodi, S. M. Khaled, Iman K. Yazdi, Michael Evangelopoulos, Naama E. Toledano Furman, Xin Wang, Federico Urzi, Sarah Hmaidan, Kelly A. Hartman and Ennio Tasciotti Department of Nanomedicine, Houston Methodist Research Institute, Houston, TX, USA

Synonyms

Stimulus-responsive polymeric hydrogels

Definition

Smart hydrogel is defined as the polymer network able to respond to external stimuli through abrupt changes in the physical nature of the network.

S

Polymer Science in Medicine

The first application of polymers in the medical field can date back to the 1940s when polymethylmethacrylate (PMMA) was used for the replacement of damaged corneas. Since then, the mechanical, physical, and chemical properties of polymers have been extensively investigated and utilized for numerous medical applications. In particular, regenerative medicine has benefited greatly from polymer research and development.

Polymers can now replace metal devices used in orthopedic settings, and the investigation into the development of new biomaterials to repair and substitute body tissues is proceeding with great momentum. Today, the use of polymers has an extensive array of applications in medicine. They have a major role in replacing damaged bones, increasing efficacy of wound repair, and fabricating external devices for dialysis, heart valves, vascular grafts, prostheses, implantable lenses and dental materials [1]. Nevertheless, polymer science still remains far from perfectly mimicking nature's ability to engineer biomacromolecule conjugates in terms of their structure, versatility, adaptability, and synthetic processes.

In 1975, Ringsdorf introduced the concept of a "pharmacologically active polymer," specifically referring to the chemical conjugation between a polymer and a drug, which represented a revolutionary approach in drug delivery [2]. Abuchowski et al. were the first to show the conjugation of PEG (polyethylene glycol) to a protein (bovine serum albumin) in 1977 [3]. The resulting conjugate was found to be less immunogenic as the PEG exerted its shielding effects on the incorporated protein. Since then, a variety of polymer-biomacromolecule conjugates have been developed and applied in a wide range of areas including bioseparation, drug/siRNA delivery, enzymatic catalytic processes, diagnostics and biosensing, and cell culture processes. Effective synthesis techniques are based on free radical polymerization. A more advanced method is the living radical polymerization (LRP), recently termed as reversible-deactivation radical polymerization [4]. This radical polymerization is aimed to reduce termination reactions (in which two active chains react with each other to become inactive), thus providing the ability to control the final molecular weight and molecular weight distribution. It also permits polymer chains to be extended as needed with the addition of monomers once the dedicated feed is done. In this respect, LRP is highly useful in producing functional "smart" polymers, as will be discussed later [5].

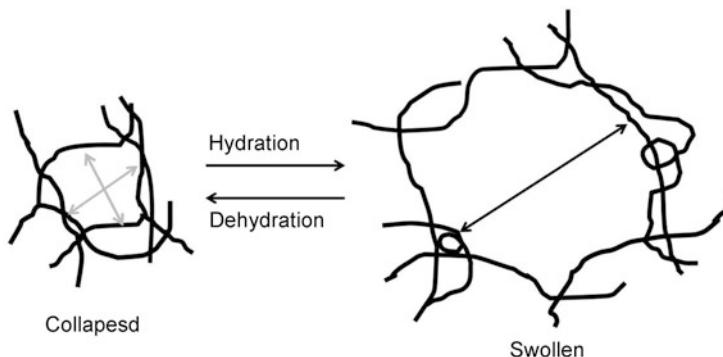
Construction of responsive polymer bioconjugates is usually achieved by one of

three methods: "grafting-to," "grafting-from," and "grafting-through." The most common method is "grafting-to" which involves the covalent or noncovalent attachment of a preformed polymer to a biomacromolecule by reactive coupling or affinity [6]. "Grafting- from" is a more recent method in which the biomacromolecule is functionalized with a moiety capable of initiating polymerization of monomers in solution. Lastly, "grafting-through" involves the attachment of the biomacromolecule to the monomer before polymerization, the generated conjugates potentially having multiple biological species attached along the polymer backbones [7].

Another aspect to be considered is the bioproperties of the polymeric chain itself. Biodegradable polymers were the first type of polymers intended for drug delivery because of the wide range of payloads that can be easily loaded and released during the degradation of the polymers. This class of polymers has been extensively studied over the last decades with successful examples such as calcium alginate and PEG-based polymer networks [8]. In addition, the advancement of nanomedicine and of its biomedical applications sparked a new breakthrough in the use of polymers for drug delivery. This application of polymers to drug delivery has been classified into five main categories: polymeric drugs, polymer-drug conjugates, polymer-protein conjugates, polymeric micelles, and polynucleic acids [9]. The improvements in synthesis, surface modifications, and characterization techniques have transformed polymers into ideal partners for many of the nanoparticles currently used in drug delivery. The fusion of these two sciences enables nanovectors to exhibit enhanced bioavailability and specificity with respect to the targeted biological sites.

Introduction of Hydrogel in Medical Field

In order to be used in the medical field, polymers must meet strict biocompatibility criteria. They have to exhibit low to no toxic effects, be hydrosoluble, dissolve into safe by-products in



Smart Hydrogels, Fig. 1 Swelling of hydrogel. *Left:* Hydrogel in the unswollen or dry state. The polymer chains are in close proximity and may interact with each other. However, when fluid enters the hydrogel, the polymer chains undergo hydration that leads to swelling due to this hydrophobic or electrostatic interaction. *Light gray*

arrows indicate pressure being exerted on the system that leads the swelling. *Right:* The hydrogel in the swollen or hydrated state where the polymer chains are fully extended and only the crosslinks prevent the material from dissolution

aqueous environments, and be nonimmunogenic to avoid activating the inflammatory response [10]. Hydrogels, also defined as hydrophilic three-dimensional polymeric networks, represent a promising option in this regard. Featured with its biocompatibility to the surrounding environment and its biodegradability, hydrogels allow for the introduction of new aspects and possibilities in the medical field. The surface of hydrogels can be reconstructed and engineered into scaffolds for the growth of synthetic tissues (pancreas, cornea, skin, bones) [11]. These materials are advanced in regenerative medicine because of their improved tolerance in the recipients for implantation and their suitability for soluble injection.

Hydrogels also provide a venue for controlled and localized drug loading, transportation, and release [12]. Recently, the applications of hydrogels have been expanded. For example, a hydrogel with embedded target biomolecules can be used as biosensor for small-scale analyte detection in diagnosis. They can be reprogrammed into tunable or self-healing coatings to interact with different microenvironments. They can also provide the opportunity for dynamic control of permeation of chemicals and drugs or serve as carriers for catalytic nanoparticles and enzymes in chemical and biochemical catalysis. Finally, hydrogels can be actuated to mimic the action of muscle in the field of biomimetic medicine.

Composed of hydrophilic cross-linked polymers, hydrogels do not dissolve in water and can hold a large amount of water or biological fluid while maintaining their structure. The ability of hydrogels to absorb water arises from the hydrophilic functional groups attached to their polymeric backbone. Their resistance to dissolution arises from cross-linking between the network chains. Figure 1 represents the polymeric network structure of a typical hydrogel interacting with the aqueous phase. While in a dehydrated or deswollen state, its polymer chains are in close proximity to each other, leaving little room for the diffusion of molecules. As the material swells, the polymer chains separate to an extent determined by the properties of the solvent. During the swelling process, the polymer chains of the hydrogel extend, decreasing the interaction that takes place between them. In this state, the swelling of the hydrogel is counteracted by the cross-linkers present within the hydrogel matrix. At its maximum hydration, the diffusion of small molecules (e.g., drugs) approaches the diffusion coefficient in pure fluid.

S

Smart Hydrogels

Hydrogels are characterized by a dynamic balance with water or other biological fluids that drives the absorption and release of drugs in the desired

environment. The presence of water in the hydrogel matrix determines the physical and chemical properties of these polymeric structures in the drug delivery field and increases their biocompatibility. Hydrogels represent a very interesting tool in regenerative medicine because they reduce the probability of inflammatory responses due to mechanical frictions with the surrounding tissues in implant surgeries. In addition, these matrices can be formulated in injectable solutions, which represent an appreciable advantage over the conventional surgical procedures.

During the last decades, hydrogels gained increasing interest in medical field because they can be designed or tailored to undergo discrete or continuous volume transformation in response to infinitesimal changes of environmental stimuli such as pH, temperature, electric field, solvent composition, salt concentrations/ionic strength, light/photon, pressure, and coupled magnetic or electric fields. Better known as smart or stimuli-responsive hydrogels, these highly maneuverable and adaptive polymers can sense changes in the environment and respond by inducing structural changes (increasing or decreasing their degree of swelling) without requiring an external driving force. This “volume-changing” phenomenon is particularly useful in drug delivery applications as drug release can be triggered upon these environmental changes.

Smart hydrogels are ideal candidates for the development of self-regulated drug delivery systems with enhanced therapeutic efficacy. Temperature and pH are the most commonly used stimuli to trigger the hydrogel’s curative action because they have biological and physiological relevance [13]. The ability for hydrogels to swell as a result of external environmental factors relies largely on three forces: (1) polymer and solvent interaction; (2) polymer elasticity; and (3) ion osmotic force, commonly referred to as electrostatic forces [14]. In the first case, the interaction with the solvent results in swelling due predominantly to the increase in translational entropy, a result of the solvent’s molecules invading the polymer network. The second force, polymer elasticity, is determined by the elastically active chains intertwining the crosslinks within the polymer

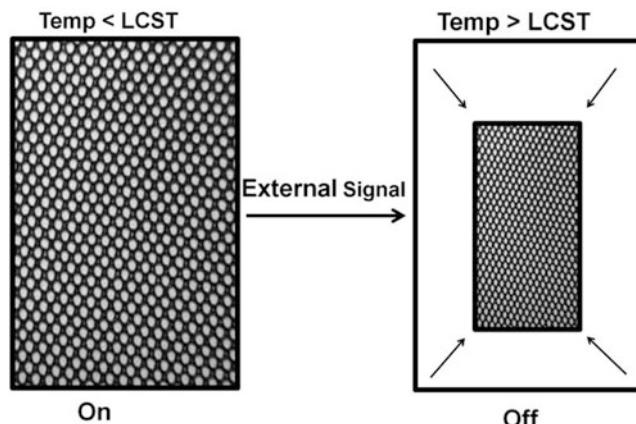
network. Specifically, this is determined by the mean number of chains per crosslink and the crosslink density increasing with temperature. While early models developed by Flory and Rehner were capable of predicting polymer swelling based on these two forces, a third force requires additional tuning [15]. In the presence of ionized groups, the third force (i.e., ion osmotic force or electrostatic force) must be considered. Similar to the Gibbs-Donnan effect, in order for electroneutrality to be enforced everywhere except at the polymer interface, counterions must be drawn into the hydrogel while cations are simultaneously released.

Temperature-Sensitive Hydrogels

Temperature-sensitive hydrogels are probably the most universally studied class of environmentally sensitive polymer systems in drug delivery research. These hydrogels are able to swell or collapse as a result of temperature fluctuation in surrounding fluid and are classified into negatively and positively thermosensitive hydrogels. Thermoresponsive hydrogels are composed of polymer chains that possess either moderately hydrophobic groups, such as methyl, ethyl, and propyl, or a mixture of hydrophilic and hydrophobic segments. Negative temperature-sensitive hydrogels are characterized by their lower critical solution temperature (LCST). Below the LCST, the hydrogel swells in solution, and above this temperature the polymer contracts. On the other hand, positive temperature-sensitive hydrogels are characterized by their upper critical solution temperature (UCST). Below UCST, the polymer shrinks, and above UCST it swells. As a result of adjusting the critical solution temperature within a physiological range, these hydrogels acquire intriguing biological activities that make them reliable candidates for drug delivery in environmental conditions in which the temperature varies. Polymeric materials such as poly(*N*-isopropyl acrylamide (PNIPAM) or poly(γ -2-(2-methoxyethoxy)-ethoxyethoxy- ϵ -caprolactone)-b-poly(γ -octyloxy- ϵ -caprolactone) have been preferred as block copolymers for

Smart Hydrogels,

Fig. 2 Schematic illustration of the “on-off” release from a squeezing hydrogel device for drug delivery (Adapted from Ref. [17])



thermosensitive nanocarriers demonstrating marked transition temperatures, allowing improved drug release at low hyperthermia. It can also occur during a brief temperature decrease (also called cold shock or cryotherapy). In this case, a thermally reversible swelling or deswelling of the block copolymers such as Pluronic F127–polyethyleneimine (PEI) leads to diffusion of the encapsulated small interfering RNA (siRNA) delivery into the cytosol [16].

Under the critical point of negative temperature-sensitive hydrogels, hydrogen bonding between hydrophilic segments of the polymer chain and water molecules dominates, leading to enhanced dissolution in water. As the temperature increases, however, interactions among hydrophobic segments are strengthened, while hydrogen bonding weakens. The net result is deswelling of the hydrogels due to interpolymer chain association through hydrophobic interactions [13]. The LCST can be changed by adjusting the ratio of hydrophilic and hydrophobic segments in the hydrogel system. It can also be done by developing copolymers of hydrophobic (e.g., *N*-isopropylacrylamide (NIPAAm)) and hydrophilic (e.g., acrylic acid (AA)) monomers. Hydrogels with negative temperature sensitivity have the potential to show an “on-off” drug release with “on” at low temperature and “off” at high temperature allowing pulsatile drug release. Figure 2 shows a schematic illustration of the on-off release [17].

Polymers having LCST below human body temperature have the potential for injectable

applications and are often combined with polysaccharides such as chitosan, alginate, cellulose, and dextran [18]. This can improve the material’s properties in terms of biocompatibility, swelling ratio, pH-sensitivity, swelling dynamics (swelling and deswelling rate), modulation of LCST, and thermal stability. Many injectable hydrogels are based on this particular feature as they remain in the Newtonian fluid form at room temperature and shrink once injected into the patient’s body. This allows for a prolonged delivery of loaded drugs, antineoplastic or analgesic drugs being just a few examples of the many possibilities. For example, Poly(*N*-isopropylacrylamide) (PNIPAAm)-based hydrogels have a great potential for biomedical applications, especially for *in vivo drug delivery*. PNIPAAm’s LCST (32°C) is in the physiological range and above the temperature of an operating room (e.g., body temperature). These hydrogels can be obtained with a wide range of crosslinking methods, and the desired features for the polymer can be easily tuned by modifying the chemical formulation of the hydrogel. For example, the LCST of PNIPAAm can be changed by adding different comonomers which will increase LCST and vice versa.

On the other hand, this polymer is not fully biocompatible as it could activate a platelet coagulation cascade. In order to overcome this limitation, the grafting of PNIPAAm with PEG (triblock copolymer) allows for the formulation of a biocompatible and biodegradable hydrogel to deliver the drug *in situ* for several hours. In addition, the

features of this copolymer can be tuned in function of the molecular weight of the PEG (up to a limit of MW 10,000). This parameter can be changed in order to modify the porosity of the hydrogel, where a higher MW causes the hydrogel to have a limited number of bigger pores because the high MW of the PEG limits the mobility of PNIPAAm [19].

Another copolymer worth mentioning is the triblock PLGA-PEG-PLGA, which showed thermosensitive biodegradability and biocompatible properties. This copolymer has been used to deliver a wide range of drugs such as insulin, calcitonin, porcine growth hormone, and testosterone, and its use can be applicable to the delivery of hydrophilic and hydrophobic molecules. The release of the drug in the body is due to both drug diffusion and the degradation of the PLGA releasing the incorporated molecules, where hydrophobic molecules are mainly released by the degradation of the copolymer and are “delivered” over a longer temporal window. The features of this class of polymers can be modified by changing some variables such as the molecular weight of PEG and PLGA, as well as the LA/GA ratio. The applications of these kinds of hydrogels range from drug delivery (cancerous, ocular, arteriovenous) to tissue engineering. Another issue of the thermoresponsive polymers is represented by temperature-sensitive peptides often associated with the widely investigated Elastin-like Polypeptide (ELP) polymers. This polymer is composed of pentapeptide monomers, derived from human tropoelastin [20], which tend to aggregate in a reversible way under heating, entrapping the payload. This biopolymer showed extraordinary biological properties as it is fully biocompatible, completely biodegradable without any immunogenic reaction, and it appeared to be ideal for the delivery of protein-based treatments. The degradation of the polymer is constant and slow when compared to the free form of the monomer.

The temperature-sensitive peptides represented a breakthrough in polymer science in the biomedical field. The exploitation of standard recombinant DNA techniques via genetic engineering allowed them to be easily synthesized

and modified in the primary sequence. This changed the chemical, physical, and biological properties of the polymer while permitting a controlled design of the peptide, according to the nature of the cargo. Furthermore, they could easily be attached to a protein-based payload during the synthesis process.

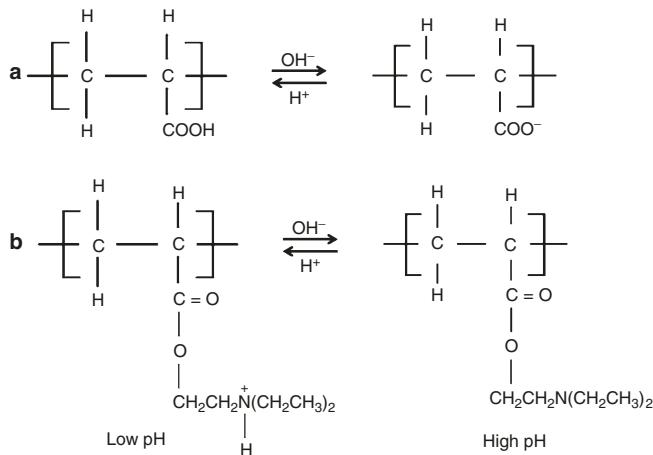
The two main features of ELP polymers are directionality and reversibility. Directionality describes the conformational state of the polymer under a specific stimulus. The ability of the peptide to associate or disassociate under heating is essential if the treatment requires rapid or prolonged drug release. Reversibility indicates the capacity of the polymer to return to its initial state after the thermal stimulus [13]. A reversible polymer could be very useful for concentrating the payload in a region of the body and obtaining a targeted release of the drug after heating through an external stimulus. Conversely, an irreversible polymer could allow drug delivery in a controlled fashion through slow degradation. This category of thermosensitive polymers is emerging as a new opportunity in cancer cell therapy [13], and future research aims to increase and optimize the biocompatibility and the tenability of these delivery systems. The challenge in the design of these thermoresponsive polymeric delivery systems lies in the use of materials that are both safe and sensitive enough to respond to slight temperature changes around the physiological temperature.

pH-Sensitive Hydrogels

Another class of smart hydrogels undergoes a volume transition with the variation of the environmental pH. pH variations have been exploited to control the delivery of molecules to intracellular compartments (such as endosomes or lysosomes) or specific organs (such as the gastrointestinal tract or the vagina) to trigger the release when subtle environmental changes are associated with pathological conditions such as cancer or inflammation. Polyacid or polybase hydrogels with ionizable groups and the hydrogels with acid-sensitive bonds whose cleavage enables the release of molecules anchored at

Smart Hydrogels,

Fig. 3 pH-dependent ionization of hydrogels. (a) Poly (acrylic acid) and (b) poly (*N,N*-diethylaminoethyl methacrylate)



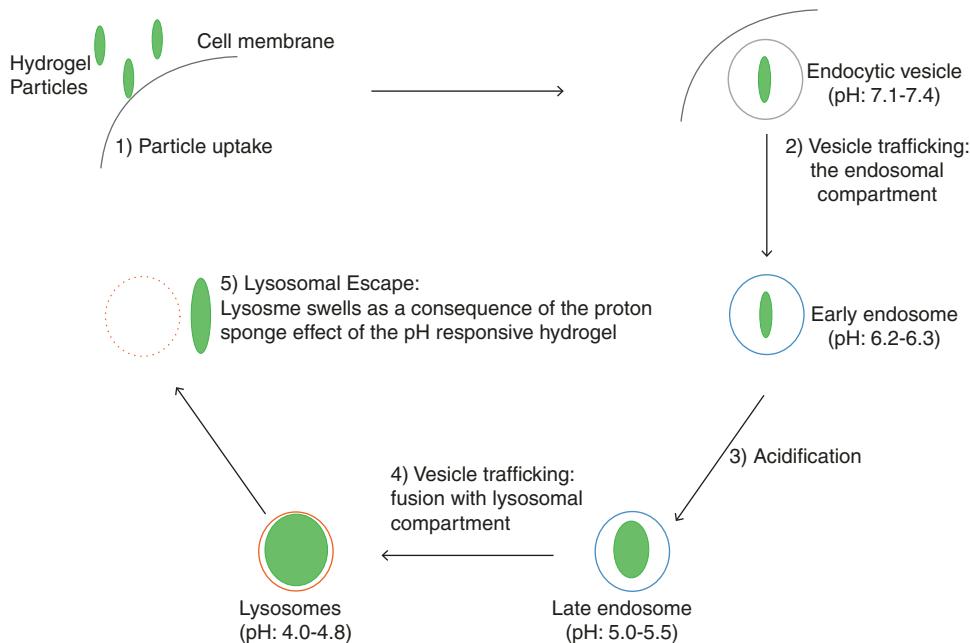
the backbone are the two strategies used to utilize response to environmental pH variation [21].

These hydrogels are characterized by the presence of ionic moieties on the polymeric backbone. These ions either accept or release protons in response to pH changes. Weak acidic groups such as acrylic and cationic acids or weak basic groups such as amines provide pH sensitivity to the polymer chain. Swelling of hydrogels sharply changes in the vicinity of the pKA or pKB values of the acidic or basic functional groups. Carboxylic pendant groups accept hydrogen at low pH but exchange it for other cations above the pKA value, and become ionized at higher pH. The hydrodynamic volume and swelling capability of these polymer chains increases sharply when their carboxylic groups become ionized and reach a plateau around pH 7. Amines, on the other hand, accept protons at low pH and become positively ionized at and below their pKB values. Hence, their swelling capability increases sharply in acidic solutions. Figure 3 shows structures of anionic and cationic polyelectrolytes and their pH-dependent ionization. Poly(acrylic acid) (PAA) becomes ionized at high pH, while poly(*N,N*-diethylaminoethyl methacrylate) (PDEAEM) becomes ionized at low pH.

The swelling and pH-responsiveness of polyelectrolyte hydrogels can be adjusted by using neutral comonomers, such as 2-hydroxyethyl methacrylate, methyl methacrylate, and maleic anhydride [13]. Different comonomers provide

different hydrophobicity to the polymer chain, leading to different pH-sensitive behavior. pH-responsive polymers were first applied as oral drug delivery systems due to the variation of the pH that characterizes the different areas of the gastrointestinal tract. Several anionic polymers such as methyl acrylic acid, methyl methacrylate, and hydroxypropyl methylcellulose phthalate have been commercially used as enteric coatings for oral delivery of protein-based drugs. In the acidic environment of the stomach, the carboxylic acid groups of the aforementioned polymers are unionized, and the particle retains its therapeutic cargo. They were shown to release the encapsulated drug molecules into the small intestine in response to the alkaline pH, because the carboxylic acid group becomes ionized and the particle swells. PNIPAAm can be classified as a thermal and pH-responsive polymer. It is used for oral drug delivery applications, exploiting its degradation characteristics in the alkaline environment of the intestine. It was shown that the polymer synthesized in combination with butylmethacrylate (BMA) and acrylic acid (AAc) is very stable at low pH, while resulting in full degradation and release of payload within the high pH environment [22].

Vesicles derived from endocytosis fuse their membranes with vesicles of the early endosomal compartment. Lysosome formation follows after several other membrane fusion steps (with late endosomal and lysosomal vesicles). During



Smart Hydrogels, Fig. 4 Schematic of the lysosomal escape. Upon cell internalization by endocytosis, hydrogel particles are trapped in the endolysosomal compartment. During their transition from early to late endosomes and

these membrane fusions inside the cytoplasm, the vesicles are proton rich, decreasing their internal pH through an active import of H^+ . This proton enrichment is permitted by an ATPase that works using ATP to pump in hydrogen ions and a negative counterion (Cl^-) to satisfy electroneutrality. Due to the presence of many ionizable groups in the backbone, a pH-responsive polymer is usually able to work as a “proton sponge” with a buffering effect on the surrounding environment. Therefore, the ATPase continues to work with the result of increasing the concentration of the counterion and affecting the osmotic balance of the vesicles. The final result of this process is the breaking of the vesicle’s membrane, caused by an increase in osmotic pressure.

Many polymers increase their volume when subjected to an acidic environment. This produces a mechanical stress on the lysosomal membranes, affecting the integrity of these organelles. Polymers that exhibit this behavior show a proton-sponge effect, and lysosomal escape is very useful for nonviral gene delivery and could potentially be

finally to lysosomes, particles undergo increased swelling due to protonation. The proton sponge effect they exert eventually leads to the breakage of the lysosomal vesicle

more efficient and less toxic than standard virus-based carriers. These molecules are usually cationic and therefore likely to be efficient in forming macrocomplexes with DNA (polyplexes) or creating new therapies based on RNA (e.g., siRNA, miRNA). Figure 4 represents a schematic of lysosomal escape performed by a pH-responsive cationic hydrogel. The polymer in this case plays three roles: (1) to physically carry the plasmid, (2) to shield the DNA payload from degradation, and (3) to perform lysosomal escape and facilitate the nuclear delivery of the cargo. Poly(ethylene imine) (PEI), poly(amidoamine) (PAMAM), Poly(*N,N*-dimethylaminoethylmethacrylate) (PDEAEM), poly(L-Lysine), and modified chitosan represent a few examples of all the polymers studied for this application.

Thermo- and pH-responsive hydrogels allow (with or without the use of copolymers) the loading and the triggered release of many drugs for the treatment of different pathological states. Table 1 reports some examples of these polymer-drug formulations and their potential clinical applications.

Smart Hydrogels, Table 1 Some applications of temperature and pH-responsive polymer

Polymer	Environmental stimulus	Copolymer	Loaded drug	Application
PNIPAAm	Temperature	Chitosan	Diclofenac	Anti-inflammatory
Glycidyl methacrylate	Temperature	Chitosan	Doxorubicin	Chemotherapeutic
β-glycerophosphate	Temperature	Chitosan	Chitosan	Antibiotic
Poly(propylene)-g-AA-g	Temperature	NIPAAm/ Chitosan	Chitosan	Antibiotic
Triblock polymer	Temperature	Poly(lactide-co-glycolide, polylactic acid, poly(ethylene glycol))	Adriamycin	Chemotherapeutic
PNIPAAm	Temperature	Alginate	Indomethacin	Anti-inflammatory
PNIPAAm	pH	Acrylic acid	Growth factors	Wound repair
Poly(ethylene imine)	pH		Nucleic acid	Gene delivery
poly(amidoamine)	pH		Nucleic acid	Gene delivery
Poly (<i>N,N</i>-dimethylaminoethylmethacrylate)	pH		Nucleic acid	Gene delivery
poly (L-Lysine)	pH		Nucleic acid	Gene delivery

Other Responsive Compounds

Recently, electroresponsive hydrogels stimulated the curiosity of the scientific community because they can unswell or bend, depending on the shape and orientation of the gel along an electric field. The gel bends when it is parallel to the electrodes, whereas unswelling occurs when the hydrogel lies perpendicular to them. This kind of polymer is usually based on a polyelectrolyte matrix (polymers that contain a relatively high concentration of ionizable groups along the backbone chain) so these polymers are also pH responsive. Weak electric fields can be used to achieve pulsed or sustained release of drug molecules through various mechanisms. For example, polypyrrole hydrogel exhibited tailored drug release as a result of a synergistic process of electrochemical reduction–oxidation and electric-field-driven movement of drug molecules [23].

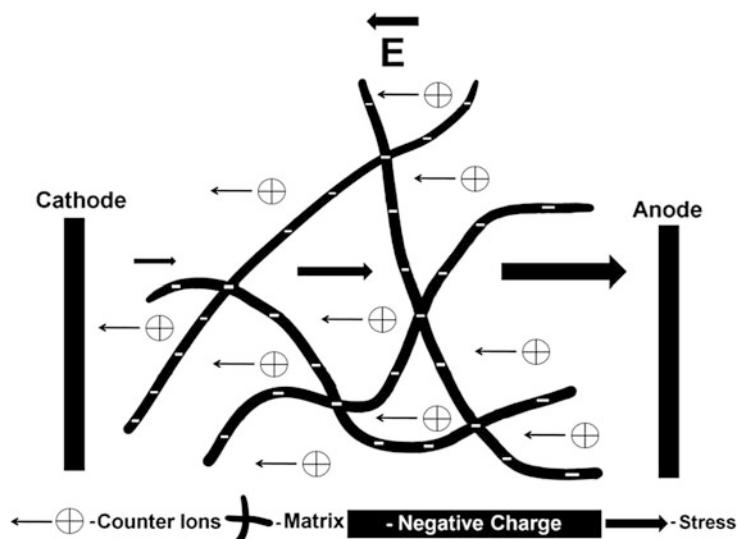
The electric field generates an interaction between the mobile counterions and the immobile charged groups of the gel's polymeric network in order to induce the phase transition. For example, in partially hydrolyzed polyacrylamide gels, the mobile H⁺ ions migrate toward the cathode while

the negatively charged immobile acrylate groups in the polymer networks are attracted toward the anode (Fig. 5). Thus, the anionic polymeric gel network is pulled toward the anode. This pull creates a uniaxial stress along the gel axis, being maximal at the anode and minimal at the cathode. This stress gradient contributes to the gel deformation, and when triggered by an electric field, these hydrogels undergo the swelling and contraction necessary to release the payload. Electroresponsive polymers represent one of the most complex challenges in biomedical polymer synthesis. Even if the electric stimulus can be finely tuned in terms of strength, duration, and impulse frequency, the use of these techniques is limited by objective safety restrictions. This kind of hydrogel was conceived to obtain a pulsatile delivery of hormones (e.g., insulin, hydrocortisone), proteins and peptides, and glucose. A few clinical applications for electroresponsive polymers are currently being investigated for dermal and transdermal drug delivery [24].

Ultrasound-triggered hydrogel-based delivery system represents a noninvasive effective approach to achieve spatiotemporal control of drug release at the desired site while preventing

Smart Hydrogels,

Fig. 5 The effect of an electric field on a polyelectrolyte gel. Positively charged counterions migrate toward the cathode, while the immobile polymeric anionic groups are attracted toward the anodes (Adapted from Ref. [34])



any possible side effects to surrounding tissues with flexibility to regulate the tissue penetration depth by tuning frequency, duty cycles, and time of exposure. Ultrasound waves trigger the release of the drug molecule through the thermal and/or mechanical effects generated by cavitation phenomena or radiation forces [25]. Combination of thermosensitive hydrogels with high-intensity focused ultrasound waves enables triggered drug release to obtain high delivery efficiency with only a mild temperature increase.

Stimulus-Sensitive Connection

This approach allows not only for the development of drug or gene delivery tools but also adaptive structures, self-healing materials, and sensing detectors that are responsive to the composition of the biological microenvironment [26]. These materials are often designed to achieve additive conformational changes of the individual subunits producing a coherent mechanical response to an external stimulus.

Stimuli can include heat, stress, pressure, electrical current/voltage, magnetic field, pH change in solvent, water, or moisture, light (photoresponsive materials), and ultrasound-responsive materials [27]. In the case of biomacromolecule conjugates, the stimuli may

confer the responsive nature of the following polymer by the ability to perform fine tuning of the solubility, stability, and/or bioactivity of the resulting conjugate. This new generation of polymer bioconjugates exhibits properties and activities that can be controlled and tuned rather than simply enhanced [28].

In this case, the polymer's function is limited to being a carrier for the cargo and targeting the area with the correct environmental conditions for drug release. Polymers such as *N*-(2-hydroxypropyl) methacrylate (HPMA), already exploited for steric stabilization of bioactive proteins, were used to coat biological or synthetic particles to provide protection against antibodies and complement factors. HPMA was also used to bind doxorubicin via a pH-sensitive linker. This polymer-bound drug showed significant improvement in the ability to kill MCF-7 breast cancer cells when compared with the drug alone. Another example of drug delivery system based on a peptide linker (Pro-Val-Gly-Leu-Ile-Gly) was composed of a chemotherapeutic agent and a dextran-based polymer. This linker is sensitive to the action of the overexpressed cancer metalloproteinases 2 and 9, while remaining stable in blood circulation. Therefore, the drug release is more concentrated at the site of the tumor where these enzymes are usually overexpressed [29].

A new frontier of material science is aiming to develop self-healing materials with an elongated lifespan through the formation of supramolecular hydrogels designed in a host-guest interaction pattern. The selective complementary interactions between host and guest molecules are versatile and can be used for the preparation of self-healing/self-repairing polymers. They can be used as both architectural materials and external coatings. The sol–gel phase transition can occur through different stimuli such as heat, pH, light, or redox environment. Multipoint crosslinks play an important role in forming the supramolecular hydrogel: usually their structure is formed by a main chain containing host molecules with a sufficient length and an appropriate number of guest molecules. For example, Cyclodextrin (CD) is used as a host molecule because it has a diverse set of applications while ferrocene (Fc) is used as a guest molecule. Polyacrylic acid (pAA) possessing CD and pAA possessing Fc are mixed together to form a supramolecular hydrogel. Fc derivatives have redox-responsive properties and variations and could induce a reversible sol–gel phase transition in a supramolecular hydrogel, resulting in a hydrogel with self-healing properties such as readhesion between cut surfaces. There are several research groups attempting to use them for a variety of different applications such as artificial molecular muscles, drug delivery systems, and vascular embolization.

Recently, stimuli-responsive polymers exhibited an ability to change their shape in reaction to a particular right stimulus. When the shape change is accompanied by the stimulus, this is the shape change material (SCM) [28]. Elastic rubber, electroactive polymer, piezoelectric material, and liquid crystal are few examples of such SCMs. Other types of stimuli-responsive materials demonstrating the ability to actively changing shape can be differentiated into shape memory materials (SMM). Although massive progress has been made in the past several years, SCMs and SMMs are starting to show their potential for different bio-based applications, although the research in this area is far from maturity.

Sensing and Signal Transduction by Swelling and Deswelling

As previously discussed, hydrogels have the potential to provide many beneficial properties in the field of drug delivery. For example, by sensing changes in hydrogel volume or weight, it is possible to measure enzymatic conversions or binding of solutes among other factors. For instance, this can be particularly beneficial in monitoring glucose levels in the tear film of the eye. This is accomplished by placing polymerized crystalline colloid array hydrogels in the front of the eye.

Unfortunately, measuring the changes in volume of such a small hydrogel can pose some challenges. One solution to this issue is to place a small hydrogel onto the tip of a fiber optic that is connected to a light source. This alternative method is advantageous because of its retention and ability to accurately measure blood metabolites. Other techniques devised by Peppas et al. and Siegel et al. involve the use of a micromachined cantilever beam with pH-sensitive hydrogel polymerized either at the top or base of the beam, measuring the deformation of the beam that occurs upon swelling of the hydrogel [8, 30].

Detection of small molecules through implantable sensors requires the consideration of additional factors such as exclusion of large molecules and isolation from the host tissue. Previously, this was accomplished by confining the hydrogel between a rigid, semipermeable membrane and a transducer. By measuring the pressure buildup from the resulting confinement, it was possible to distinguish analyte-induced changes within the hydrogel. Other techniques have been employed to measure changes in pH, glucose levels, and ionic strength [31]. Nevertheless, the confinement of a hydrogel into a tight space is not always easy, and complications can occur. For example, if the hydrogel is too small, the empty space could make it difficult to measure the pressure changes that occur upon swelling. Conversely, stress imposed when the hydrogel is too large can also make it difficult to measure pressure fluctuations. To overcome this issue, one group embedded superparamagnetic iron oxide nanoparticles into the hydrogel network [32]. After lamination onto a planar coil,

the swelling of the hydrogel can be measured by quantifying the magnetic permeability following stimulation with radio frequency.

Although many beneficial measurements can be obtained by evaluating fluctuations in weight and volume, strategically using the swelling and deswelling effects for drug release can also prove useful. Specifically, the expansion in mesh size resulting from swelling has previously been shown to provide favorable drug release properties [33]. Similarly, strong contractions in the hydrogel network can also provide increased release properties. Together, these features have resulted in a multitude of delivery vectors and continue to transform the field of nanotechnology.

Summary

Smart hydrogels, a special class of polymer networks, have offered new breakthroughs in the medical field due to their exceptional properties, engineering flexibility, natural abundance, and ease of manufacturing. This entry represents a basic overview of classifications and material properties of smart hydrogels for guided drug delivery and applications in regenerative medicine. While providing a brief representation on the current trends and advancement of polymer science in medicine, we have also revealed some cutting-edge progress in the field originating from the blending of nanotechnology with polymer science.

Cross-References

- [Drug Delivery](#)
- [Nanocomposites](#)
- [Nanomedicine](#)
- [Nanoparticles](#)

References

1. Navarro, M., Michiardi, A., Castano, O., Planell, J.: Biomaterials in orthopaedics. *J. R. Soc. Interface* **5**, 1137–1158 (2008)
2. Ringsdorf, H.: Structure and properties of pharmacologically active polymers. *J. Polym. Sci. Polym. Symp.* **51**, 135–153 (1975)
3. Abuchowski, A., McCoy, J.R., Palczuk, N.C., van Es, T., Davis, F.F.: Effect of covalent attachment of polyethylene glycol on immunogenicity and circulating life of bovine liver catalase. *J. Biol. Chem.* **252**, 3582–3586 (1977)
4. Jenkins, A.D., Jones, R.G., Moad, G.: Terminology for reversible-deactivation radical polymerization previously called “controlled” radical or “living” radical polymerization (IUPAC recommendations 2010). *Pure Appl. Chem.* **82**, 483–491 (2009)
5. Cobo, I., Li, M., Sumerlin, B.S., Perrier, S.: Smart hybrid materials by conjugation of responsive polymers to biomacromolecules. *Nat. Mater.* **14**, 143–159 (2015)
6. Lutz, J.-F., Börner, H.G.: Modern trends in polymer bioconjugates design. *Prog. Polym. Sci.* **33**, 1–39 (2008)
7. Ivanov, A.E., Edink, E., Kumar, A., Galaev, I.Y., Arendsen, A.F., Bruggink, A., et al.: Conjugation of penicillin acylase with the reactive copolymer of *N*-Isopropylacrylamide: a step toward a thermosensitive industrial biocatalyst. *Biotechnol. Prog.* **19**, 1167–1175 (2003)
8. Keys, K.B., Andreopoulos, F.M., Peppas, N.A.: Poly(ethylene glycol) star polymer hydrogels. *Macromolecules* **31**, 8149–8156 (1998)
9. Schmaljohann, D.: Thermo- and pH-responsive polymers in drug delivery. *Adv. Drug Deliv. Rev.* **58**, 1655–1670 (2006)
10. Anderson, J.: The future of biomedical materials. *J. Mater. Sci. Mater. Med.* **17**, 1025–1028 (2006)
11. Alijotas-Reig, J., Garcia-Gimenez, V.: Delayed immune-mediated adverse effects related to hyaluronic acid and acrylic hydrogel dermal fillers: clinical findings, long-term follow-up and review of the literature. *J. Eur. Acad. Dermatol. Venereol.* **22**, 150–161 (2008)
12. Peppas, N.A.: Hydrogels and drug delivery. *Curr. Opin. Colloid Interface Sci.* **2**, 531–537 (1997)
13. Chilkoti, A., Dreher, M.R., Meyer, D.E., Raucher, D.: Targeted drug delivery by thermally responsive polymers. *Adv. Drug Deliv. Rev.* **54**, 613–630 (2002)
14. Erman, B., Flory, P.J.: Critical phenomena and transitions in swollen polymer networks and in linear macromolecules. *Macromolecules* **19**, 2342–2353 (1986)
15. Flory, P.J., Rehner, J.: Statistical theory of chain configuration and physical properties of high polymers. *Ann. N. Y. Acad. Sci.* **44**, 419–429 (1943)
16. Lee, S.H., Choi, S.H., Kim, S.H., Park, T.G.: Thermally sensitive cationic polymer nanocapsules for specific cytosolic delivery and efficient gene silencing of siRNA: swelling induced physical disruption of endosome by cold shock. *J. Control. Release* **125**, 25–32 (2008)
17. Gutowska, A., Seok Bark, J., Chan Kwon, I., Han Bae, Y., Cha, Y., Wan, K.S.: Squeezing hydrogels for controlled oral drug delivery. *J. Control. Release* **48**, 141–148 (1997)

18. Webber, R.E., Shull, K.R.: Strain dependence of the viscoelastic properties of alginate hydrogels. *Macromolecules* **37**, 6153–6160 (2004)
19. Alexander, A., Ajazuddin, Khan, J., Saraf, S., Saraf, S.: Polyethylene glycol (PEG)-poly(*N*-isopropylacrylamide) (PNIPAAm) based thermosensitive injectable hydrogels for biomedical applications. *Eur. J. Pharm. Biopharm.* **88**, 575–585 (2014)
20. Urry, D.W.: Physical chemistry of biological free energy transduction as demonstrated by elastic protein-based polymers. *J. Phys. Chem. B* **101**, 11007–11028 (1997)
21. Mura, S., Nicolas, J., Couvreur, P.: Stimuli-responsive nanocarriers for drug delivery. *Nat. Mater.* **12**, 991–1003 (2013)
22. Serres, A., Baudyš, M., Kim, S.: Temperature and pH-sensitive polymers for human calcitonin delivery. *Pharm. Res.* **13**, 196–201 (1996)
23. Ge, J., Neofytou, E., Cahill, T.J., Beygui, R.E., Zare, R.N.: Drug release from electric-field-responsive nanoparticles. *ACS Nano* **6**, 227–233 (2011)
24. Peppas, N.A., Bures, P., Leobandung, W., Ichikawa, H.: Hydrogels in pharmaceutical formulations. *Eur. J. Pharm. Biopharm.* **50**, 27–46 (2000)
25. Epstein-Barash, H., Orbey, G., Polat, B.E., Ewoldt, R.H., Feshitan, J., Langer, R., et al.: A microcomposite hydrogel for repeated on-demand ultrasound-triggered drug delivery. *Biomaterials* **31**, 5208–5217 (2010)
26. Yang, Y., Urban, M.W.: Self-healing polymeric materials. *Chem. Soc. Rev.* **42**, 7446–7467 (2013)
27. Sun, L., Huang, W.M., Ding, Z., Zhao, Y., Wang, C.C., Purnawali, H., et al.: Stimulus-responsive shape memory materials: a review. *Mater. Des.* **33**, 577–640 (2012)
28. Lu, H.B., Huang, W.M., Yao, Y.T.: Review of chemoresponsive shape change/memory polymers. *Pigm. Resin Technol.* **42**, 237–246 (2013)
29. Liu, W., MacKay, J.A., Dreher, M.R., Chen, M., McDaniel, J.R., Simnick, A.J., et al.: Injectable intratumoral depot of thermally responsive polypeptide–radionuclide conjugates delays tumor progression in a mouse model. *J. Control. Release* **144**, 2–9 (2010)
30. Lei, M., Ziae, B., Nuxoll, E., Iván, K., Noszticzius, Z., Siegel, R.A.: Integration of hydrogels with hard and soft microstructures. *J. Nanosci. Nanotechnol.* **7**, 780–789 (2007)
31. Lin, G., Chang, S., Hao, H., Tathireddy, P., Orthner, M., Magda, J., et al.: Osmotic swelling pressure response of smart hydrogels suitable for chronically implantable glucose sensors. *Sensors Actuators B Chem.* **144**, 332–336 (2010)
32. Song, S.H., Park, J.H., Chitnis, G., Siegel, R.A., Ziae, B.: A wireless chemical sensor featuring iron oxide nanoparticle-embedded hydrogels. *Sensors Actuators B Chem.* **193**, 925–930 (2014)
33. Lin, C.C., Metters, A.T.: Hydrogels in controlled release formulations: network design and mathematical modeling. *Adv. Drug Deliv. Rev.* **58**, 1379–1408 (2006)
34. Murdan, S.: Electro-responsive drug delivery from hydrogels. *J. Control. Release* **92**, 1–17 (2003)

Soft Actuators

- [Organic Actuators](#)
-

Soft Lithography

- [Microcontact Printing](#)
 - [Nanoscale Printing](#)
-

Soft Matter

- [Dissipative Particle Dynamics, Overview](#)
-

Soft X-Ray Lithography

- [EUV Lithography](#)
-

Soft X-Ray Microscopy

- [Selected Synchrotron Radiation Techniques](#)
-

S

Soil/Terrestrial Ecosystem/Terrestrial Compartment

- [Exposure and Toxicity of Metal and Oxide Nanoparticles to Earthworms](#)
-

Solar Cells

- [Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components](#)

Sol-Gel Method

Bakul C. Dave and Sarah B. Lockwood
Department of Chemistry and Biochemistry,
Southern Illinois University Carbondale,
Carbondale, IL, USA

Synonyms

Chemical solution deposition; Gel chemical synthesis; Silica gel processing; Wet chemical processing

Definition

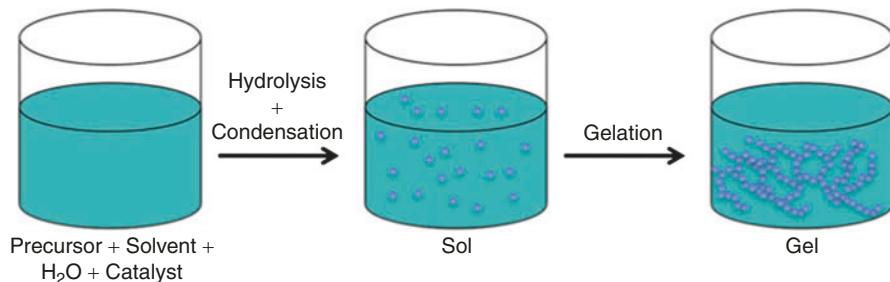
The sol-gel method is a wet chemical process of making oxide-based materials starting from hydrolyzable precursors via hydrolysis and condensation. The precursors usually contain weaker ligands as compared to water such as halides, nitrates, sulfates, alkoxides, or carboxylates. The hydrolyzed precursors then condense together to form small colloidal nanoparticles suspended in a liquid called a sol. Further polycondensation of the sol particles leads to an extended oxo-bridged network of polymeric oxide-based materials. The as-formed gels obtained by the sol-gel method are biphasic materials that contain gel network along with a significant amount of liquid phase. Drying of these gels, either under ambient conditions or at elevated temperature, can lead to expulsion of solvent phase to form dense materials. Because the method utilizes molecular precursors, it offers appealing prospects for the design of nanomaterials with a degree of control over their structure, composition, dimension, morphology, organization, geometry, and bulk architecture. While the method can be used to make bulk monoliths, it is particularly suited for fabricating nanomaterials of varied compositions in different shapes, geometries, and dimensionalities including nanoparticles, fibers, thin films, and coatings.

Overview

The sol-gel method is a solution-based method of making ceramic particles, powders, coatings, films, and monolithic objects [1]. The process is commonly used for making oxides, however, silica-based materials constitute a primary archetype system. With highly electropositive metals, the precursor molecules can react rapidly with water to undergo facile hydrolysis under ambient conditions. However, precursors based on alkoxy silanes usually require the use of catalysts, refluxing, or use of ultrasound for the reaction to proceed at an appreciable rate. The liquid sol containing nanoscale particles can be used to form powders, fibers, coatings, or monolithic forms by allowing the reaction to proceed under specific processing conditions. In addition to the formation of pure oxides, the method can also be used for making organic-inorganic hybrids and organically modified silicas as well as bioceramic hybrids [2].

In the sol-gel method, a suitable molecular precursor is hydrolyzed to generate a solid-state polymeric oxide network (Fig. 1). Initial hydrolysis of the precursor generates a liquid sol, which ultimately turns to a solid, porous gel. The gels formed this way are porous and contain a substantial amount of solvent phase. Slow drying of the gels under ambient conditions (or at elevated temperature) leads to evaporation of the solvent phase. The dried gels – termed xerogels – are substantially less porous than wet gels. Since the method begins from a molecular species, it offers a degree of control over structure and properties that can be tailored at the molecular level. The method can be used to prepare a wide range of oxides. However, the most well-known and extensively studied examples are those based on silica chemistry where there is a good understanding of how the chemical parameters can be used to control the properties of the final product.

In general, the applications of the sol-gel method can be loosely divided into three interdependent aspects of synthesis, processing, and properties, which are a direct function of the structural composition, morphology, and microstructure. The sol-gel



Sol-Gel Method, Fig. 1 Schematic depiction of the sol–gel method

method for synthesis of materials is particularly appealing because with a chemical modification of the precursor at molecular level, the functional properties of the final material can be changed. Therefore, by selectively integrating specific organic functional groups into the precursor at the molecular level, it is possible to introduce desired properties into the product sol–gel inorganic–organic hybrid material.

The sol–gel chemistry is feasible with any metal precursor that is hydrolytically unstable. Typical precursors used are alkoxides, halides, or carboxylates. With silica chemistry, alkoxides [Si(OR)₄] are predominantly used. In addition, it is possible to use organically modified alkoxides [(OR)₃SiR'] to incorporate organic functionalities in glasses. The R' group can be any organic group, including polymers or oligomers. When a precursor containing an organic group is hydrolyzed, it leads to formation of an ORganically MOdified SILicate (ORMOSIL) glass. The organic R' groups can be simple hydrocarbons or can include other functional groups. In addition, it is possible to make glasses with precursors of the type [(OR)₃Si-R-Si[(OR)₃]], where the organic functionality is present as a spacer group. The precursors can be used alone, or alternatively, a combination of precursors can be used to make hybrid or composite materials. Finally, prior to gelation, other species such as organic-, inorganic-, and biomolecules and/or polymers can be added to the liquid sol to make composite materials. When large biomolecules are added to the glass, they usually become physically entrapped in the glass but nonetheless retain

their structure and properties and the resultant materials exhibit biological function and activities [3].

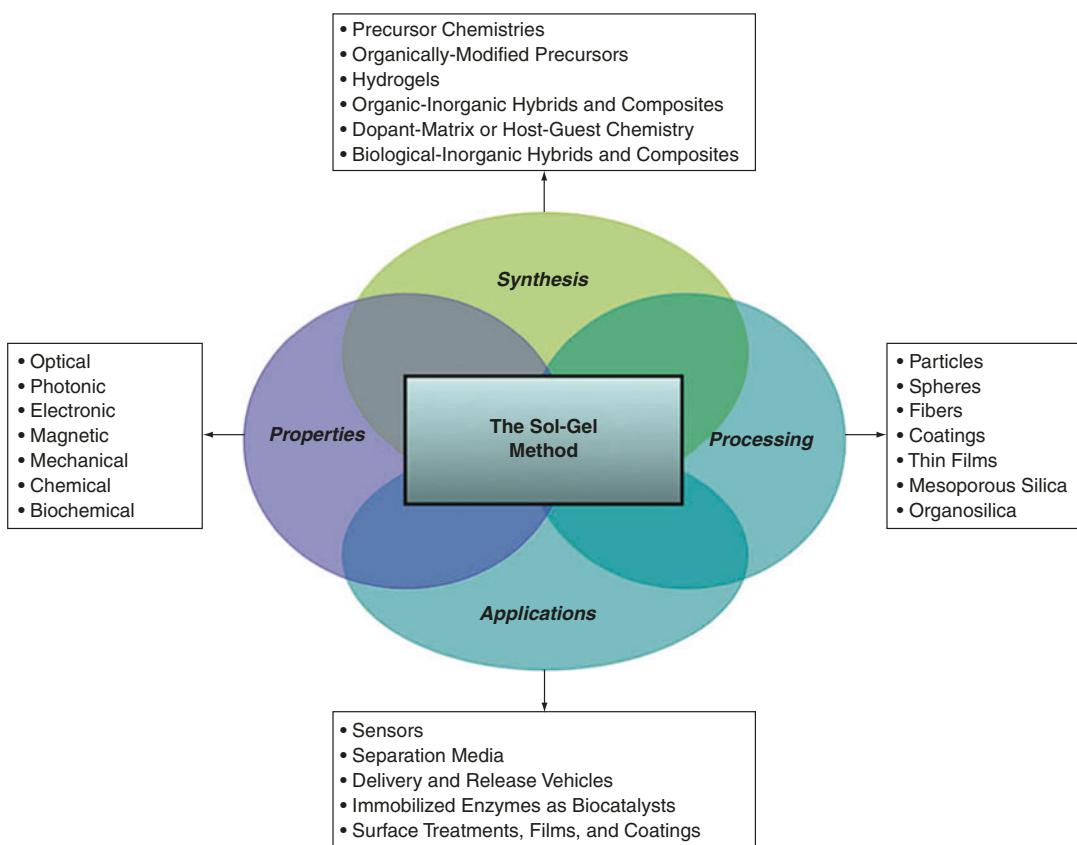
The sol–gel method is primarily a method at the interface of synthesis and processing. Synthesis refers to the making of new materials by means of specialized chemical precursors through arrangement of precise functionalities at the molecular level. On the other hand, processing refers to the specific organization of matter at the nano-, micro-, and macroscopic level. Processing of gels is greatly facilitated by the stability of the liquid sol, which enables construction of materials with good control over their nanostructure, shape, and overall geometry. The final gels can be fabricated in a variety of shapes, geometries, and configurations, including monoliths, coatings, thin films, fibers, and powders. The liquid sol can be used to make monolithic geometries whose final shape is determined by the container used. From the perspective of nanotechnology, the method provides a rather facile pathway to construction of different nanomaterials. Formation of oxide-based nanoparticles is a key area in use of the sol–gel method due to the unique opportunities provided by the solution-based method that proceeds via the sol state characterized by dispersed nanoscale particulate matter. The liquid sol can be used for dip-coating, spin-coating, and spray-coating to prepare films and coatings of varied thicknesses on diverse surfaces including glass, metals, composites, or plastics. The versatility of the method has resulted in a burgeoning utility of the sol–gel method in diverse fields.

Applications are the key drivers in the utilization of the sol-gel method in technology. Because of its versatility, the method has found widespread applications in all walks of science and engineering including, optics, photonics, microelectronics, and biotechnology to name a few. The intrinsic properties of sol-gel materials such as optical transparency, porosity, surface area along with chemical and physical stability confer a unique vantage point for their applications in different areas. Different functions and properties can be imparted to the parent glass by incorporation of additional organic, inorganic, or biological entities. Particularly important in this direction are the organosilica sol-gels and hybrid materials, which integrate the properties of both organic and inorganic materials and constitute a novel platform for development of new materials and devices from a molecular perspective. Similarly, the access to formation of bio-hybrid via integration of biological species makes the method particularly suited for the fabrication of biological-inorganic hybrids and composites. For the design of advanced materials, the advantage of using sol-gel derived glasses is that the parent silica material is structurally inert, functionally inactive, and operationally nonresponsive. Therefore, by selectively integrating specific (bio)chemical entities into the glass (either through precursor modification or through encapsulation), it is possible to introduce desired structural, functional, and operational properties in a modular fashion. The introduction of properties can be singular, sequential, or even multiple. The strategy offers a powerful approach for designing a diverse range of sol-gel-derived materials whose properties can be tailored with a degree of control over the compositional, morphological, and functional characteristics of the product materials. This flexibility enables exploring new opportunities, and materials made using the sol-gel method have found a variety of applications in optics, photonics, magnetics, biocatalysis, detection, sensing, controlled release, and separation in addition to providing access to a range of structures and morphologies in the form of particles, fibers, coatings, and other nanoscale, nanostructured, and nanoporous materials.

The sol-gel method has brought about a fundamental change in synthesis and processing of inorganic materials. The fact that final materials in different geometries and morphologies can be prepared from a solution circumvents the need to go via conventional solid-state high temperature processing pathways. The solution-based method provides considerable flexibility in making pure, homogenous materials as well as composites and hybrids. The method can be used to make materials by using a given precursor, or alternatively, by mixing a plurality of different precursors, it is possible to form multicomponent mixed-oxides. Similarly, the method can be used to make hybrids and composites via integration of polymers or other organic/biological components to obtain materials with varied properties and applications (Fig. 2). The low temperature at which the materials can be prepared makes the method economical and environmentally friendly for practical applications.

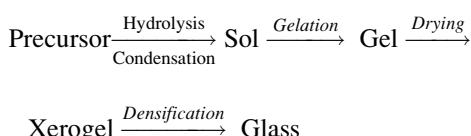
Basic Methodology

The sol-gel method of making glasses begins with the reaction of hydrolyzable precursors with water. The most commonly used precursors are alkoxides. The process is illustrated herein with the alkoxy silanes as an example; however, the reaction sequences are analogous for other metals. These precursors are dissolved in a compatible solvent such as an alcohol followed by addition of water to initiate hydrolysis. Often a catalyst is used to accelerate the reaction. The nonpolar nature of alkoxy silanes makes them immiscible with water, and therefore, the reactions proceed at very slow rates even in the presence of a catalyst. The mixing can be enhanced by use of ultrasound, or alternatively, the reaction mixture is refluxed at elevated temperature to facilitate the reaction. Under these conditions, first the alkoxides react with water and form hydroxylated species which then condense to form colloidal nanoparticles containing oxo-bridges. As the reactions proceed, the sol particles combine to form an extended network and the reaction mixture becomes viscous, ultimately leading to the formation of a



Sol-Gel Method, Fig. 2 Representative summary of significant aspects of the sol-gel method

solid gel. The as-prepared gels are biphasic porous materials with interstitial solvent phase. Drying of the gels under ambient conditions results in evaporation of a majority of the solvent phase (water and alcohol) and the gels shrink by up to 80 % in volume. These dried glasses – termed xerogels – are mechanically and dimensionally stable materials that contain some solvent phase, which can be further dried at elevated temperatures to form dense glasses. The overall method of making oxides from alkoxides is represented as



The sequence is characterized by a series of reactions along with physical and chemical changes

that take place along the precursor to sol to gel to xerogel to glass transformation steps. The properties of materials depend upon the conditions employed and by controlling the parameters under which these steps take place one can modulate, regulate, and fine-tune the properties of the materials at each stage.

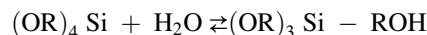
The Precursor: The nature, composition, and structure of precursor molecules is central to the properties of final product. Controlling the number of hydrolyzable versus nonhydrolyzable groups on the precursor provides an easy pathway to influence the chemical makeup of the materials obtained by the sol-gel method. Therefore, in order to obtain pure oxide-based materials, one can use precursors of the type $\text{Si}(\text{OR})_4$ wherein all the alkoxides are replaced with oxygen atoms in the product due to hydrolysis and condensation. Use of nonhydrolyzable ligands on the precursors

ensures that these groups will remain in the final material thereby altering the structural, morphological, and functional characteristics of the product. Precursors of the type $(OR)_3SiR'$ where R' is organic group provide an easy access to introduce organic functional groups that can impart specific properties. Additionally, use of multiple precursors with different functional groups provides further access to tailoring the final properties.

Alkoxides are the preferred precursors of choice due to the fact that the by-product of their hydrolysis is the corresponding alcohol, which can easily be evaporated to obtain relatively pure materials [4]. The rates of hydrolysis and condensation depend upon the nature of the alkoxide groups, which can lead to dramatic variations in the rates of the reactions as well as product morphologies. The rates of hydrolysis of alkoxides depend upon the polarity of M-OR bond as well as the steric environment around the central metal atom which can influence the approach of water molecule for hydrolysis of the bond. In general, more polar alkoxides are more susceptible to hydrolysis in a polar solvent environment. The inductive and steric effects of the alkoxide ligands can also alter the hydrolysis rates by altering the polarity of the M-OR bonds, and by modulating the access of water molecules to effect the hydrolysis step. Usually, the hydrolysis step is much slower with longer chain and/or branched alkoxides which inhibit the access of water molecules to the metal site. A catalyst is typically used, especially with silicon alkoxides, to accelerate the hydrolysis step. On the other hand, alkoxides of transition metals and other highly electropositive elements usually exhibit much faster hydrolysis due to the intrinsic polarity of the M-OR bond along with the ability of these elements to accommodate the incoming water molecules via coordination to the metal site.

Hydrolysis: The first step of the sol-gel reaction is the replacement of alkoxide groups in the precursor molecule with water. The alkoxide groups leave as the corresponding alcohols along with formation of hydroxylated silanol species. Since alcohol is a by-product of the hydrolysis reaction, carrying out the reaction in the same alcohol as the corresponding alkoxide can alter

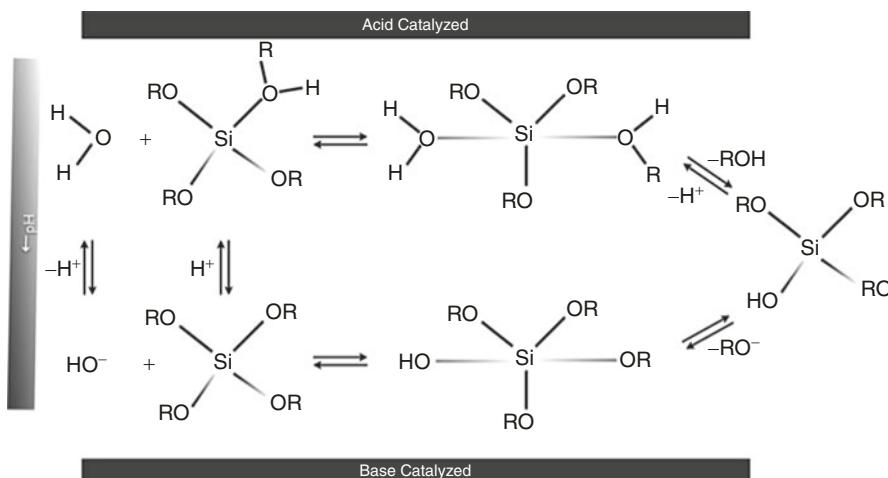
the reaction equilibrium. If a different alcohol is used as a solvent then, due to its predominance, there is usually a replacement of the alkoxide ligand via transesterification. The ease with which the reaction occurs depends upon the miscibility of the precursors with water, its hydration characteristics, and its ability to accommodate water molecules in the primary coordination sphere. The products of the hydrolysis reaction are the hydroxylated precursor and the corresponding alcohol. The general reaction is given as



The hydrolysis reaction proceeds via associative (or S_N2) pathways such that the transition state is characterized by an increase in coordination number. The reaction is usually acid or base catalyzed to make it occur at an appreciable rate. The general reaction under both conditions proceeds via associative pathways through the aid of the predominant catalytic species, namely, protons in acidic conditions and the hydroxide ions under basic conditions.

Under acidic conditions, the reaction occurs in three steps (Fig. 3). First, the alkoxide group gets protonated, which makes the metal-alkoxide bond weaker and the silicon atom electron-deficient. Second, the water molecule binds to the silicon atom to form a positively charged pentacoordinate transition state. Finally, the release of alcohol molecule completes the reaction. Under basic conditions, the reaction also occurs in three steps (Fig. 3). First, the water molecule loses its proton to form the hydroxide ion. Second, the hydroxide ion binds to the silicon atom to form a negatively charged transition state. Finally, the alkoxide ion dissociates to form the hydroxylated species.

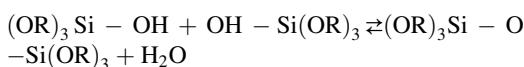
The replacement of alkoxides to form hydroxylated species proceeds at different rates under acidic and basic conditions. Under acidic conditions, subsequent hydrolysis steps are retarded after the initial substitution of alkoxide, while in basic medium, these follow-on steps are accelerated. In other words, the hydrolysis of $[(OR)_4 - nSi(OH)_n]$ (where n varies from 0 to 4) species in acidic medium is faster for smaller values of n



Sol-Gel Method, Fig. 3 Hydrolysis mechanisms in acid or base catalyzed reactions

while in a basic medium the hydrolysis is faster for larger values of n. As such, the formation of multihydroxylated species predominates in the basic medium. The rate and extent of hydrolysis can be altered by use of suitable solvents that can change the polarity of the medium, and consequently, the stability of polar versus nonpolar species to shift the equilibrium. Solvents that interact with water by hydrogen bonding interactions also alter the activity of water in the solution and affect the rate and extent of hydrolysis. Solvents that can homogenize the reactants to make them more miscible augment the rates and shift the equilibrium toward products. Finally, use of ultrasound can also facilitate mixing of the precursor with water and can accelerate the reaction.

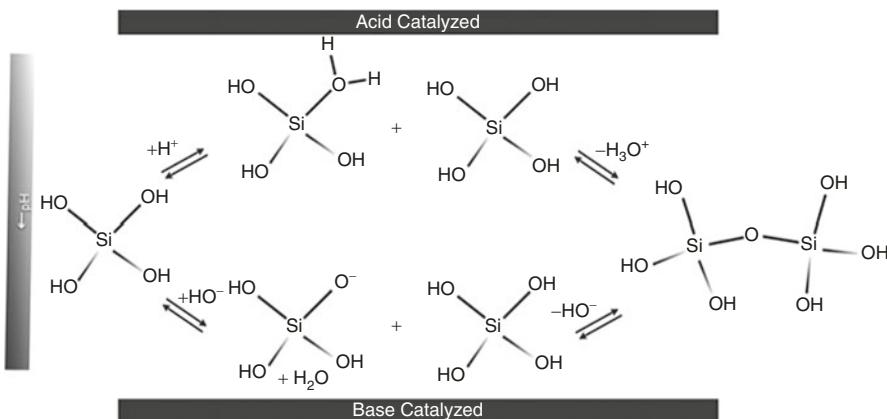
Condensation: Once the reactive hydroxylated species are formed, they begin to undergo condensation reaction to form Si-O-Si linkages to generate oligomers and polymers. The reaction can be represented as



As with hydrolysis, the condensation reactions are dependent upon pH of the medium and can be acid- or base-catalyzed (Fig. 4). Under conditions of acid catalysis, the reaction proceeds in three steps. First, there is a rapid protonation of the

silanol group to form a positively charged species. Second, the protonated positively charged species is attacked by a neutral hydroxylated species. Finally, an elimination of the hydronium ion completes the reaction. Base catalysis follows a slightly different sequence. First, there is rapid deprotonation of the silanol group to form an anionic silicate species. Second, the negatively charged species coordinates with the neutral hydroxylated species via formation of a pentacoordinate transition state. Finally, the expulsion of hydroxide ions completes the reaction sequence.

Analogous to hydrolysis reactions, the rates of the condensation reaction are dependent upon pH of the medium. Condensation reaction for $[(\text{OR})_4 - n\text{Si}(\text{OH})_n]$ (where n varies from 1 to 4) species in acidic medium is faster for smaller values of n while in a basic medium it is faster for larger values of n. As such, base-catalyzed systems, with a predominance of multihydroxylated species, are characterized by faster gelation times as compared to acid-catalyzed reactions. Silanol ($\text{Si}-\text{OH}$) groups become more acidic as the silicon atom gets increasingly surrounded by the electron withdrawing oxo groups. This high acidity of silanols with the increasing extent of condensation results in maximum condensation rates in the intermediate pH range of 3–5. In very high pH region



Sol-Gel Method, Fig. 4 Condensation mechanisms in acid or base catalyzed reactions

($\text{pH} \geq 10$), condensation reactions occur but gelation is thwarted by repulsion due to surface charges leading to formation of discrete particles. As water is a by-product of condensation reactions, the amount of water in the reaction mixture influences the equilibrium. Reaction media high in water content favor hydrolysis but hinder condensation. Similarly, more polar and hydrogen bonding solvents can stabilize the hydroxylated species to retard the condensation process. Finally, the steric effects of the remaining alkoxo groups can also influence the rate with bulkier groups hindering the condensation reactions.

Gelation: As the hydrolysis and condensation reactions continue, the reaction medium becomes more viscous due to increased degree of polymerization. The viscous sol continues to evolve until the viscosity reaches a critical point where the liquid can no longer maintain its fluidity. Gelation is characterized as the point along the hydrolysis-condensation reaction coordinate when the liquid sol turns to a solid gel. Gelation is typically defined by gelation time (t_g) as the time taken from the start of the reaction till the formation of a solid gel. Gelation times are dependent upon a variety of factors, as discussed above, that control the rate and extent hydrolysis and condensation reactions. At gelation time, the growing polymer structure essentially gets “frozen” into an immobile state. The initial gels formed have high viscosity but they lack the mechanical stability due to

lack of an extensive interconnected network. Even though the physical state of the reaction mixture changes from liquid to solid, the hydrolysis and condensation reactions still continue to evolve leading to formation of additional Si-O-Si linkages. The increasing degree of condensation results in enhanced mechanical strength over time even after gelation has occurred.

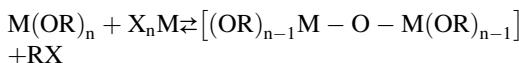
Aging: The as-formed gel is a porous structure with pendent silanols and even some remaining unhydrolyzed alkoxides on the polymeric network. Aging of the gels under closed conditions, usually extending over a period of hours to days, is employed to enable completion of these reactions. During aging, continuation of hydrolysis and condensation reactions leads to an increase in cross-linking and overall stiffening of the gel network. During aging, the gels shrink and expel the solvent phase due to a decrease in internal pore volume caused by extensive condensation of silanols to siloxane linkages. Aging is a necessary step in processing of bulk gels to ensure mechanical stability of these materials during subsequent stages.

Drying: Drying of the gels either under ambient conditions or elevated temperature is used to expel solvent from the gels. The elimination of solvent from the pores of the gels compacts the network and facilitates further condensation reactions between silanols on the surface of pores. The ambiently dried gels can lose up to 70–80 % of

the solvent and shrink considerably due to compaction of the pore network. In order to completely dry the gels, use of elevated temperatures is necessary. Drying enhances the mechanical stability of the gels, and the dried gels are considerably stiffer and stronger materials as compared to aged gels. The dried gels have a greater density and are dimensionally invariant under ambient conditions.

Sintering/Densification: Heat treatment of the gels further eliminates the porous structure to obtain ceramics or dense glasses. During heat treatment, adsorbed and hydrogen-bonded solvent is removed and the collapsing pore structure causes further condensation of silanols to siloxane linkages. Heat treatment makes the gels nonporous and harder with densities and structure approaching close to fused silica.

Nonhydrolytic sol-gel method: The sol-gel method depends upon water to initiate the hydrolysis of precursors. This presents a challenge for system that are either unstable in water or too reactive. In order to circumvent these issues, the nonhydrolytic sol-gel method [5] utilizes nonaqueous solvent medium and does not require the hydrolysis step. While there exist many different variants, the general nonhydrolytic sol-gel method can be represented as



where X = halide, $-\text{OR}'$, or $-\text{O}_2\text{CR}'$. The reaction bypasses the hydrolysis process altogether and leads to direct formation of oxo-bridged linkages. The reaction is essentially a condensation pathway that is usually carried out at elevated temperatures. The nonhydrolytic process is advantageous in a system wherein a greater degree of control over the relative rates of hydrolysis and condensation to fine tune product microstructure, porosity, and morphology may not be possible with the hydrolytic pathway.

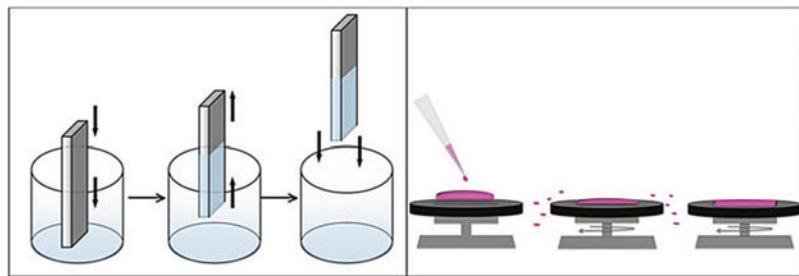
Formation of nanoparticles and powders: The molecule-nanomaterial interface along the precursor to sol reaction coordinate provides a convenient access to a variety of nanoscale systems. Powders such as precipitated or fumed silica

have been long used in commercial applications. A key challenge has been to obtain uniform monodisperse nanoscale particles with a control over their nuclearity, shape, and dimensions. The issue of uniformity and homogeneity, to a certain extent, is addressed by utilizing specific reaction conditions to prevent interparticle interactions. Two methods commonly used to stabilize isolated discrete nanoparticles are based on manipulating electrostatic and steric factors. As discussed above, highly basic conditions favor rapid hydrolysis and condensation while producing negative surface charges which favor the formation of particulate structures. The Stober method of forming particles specifically takes advantage of this approach. Uniform size spherical particles can be obtained under basic conditions by using dilute solutions to prevent interparticle aggregation and to allow dissolution-growth phenomena to occur without precipitation. Usually the concentration of the precursor in the reaction mixture is optimized to prevent interparticle polycondensation and prevent formation of larger particles due to particle growth and coalescence. Sometimes, a synthesis strategy of combined acid and base catalysis is used for the initial formation of a stable sol under acidic conditions followed by addition of base to facilitate particle growth. Optimized reaction conditions that favor dynamic surface reorganization and reconstruction to provide uniform spherical particles are commonly utilized. Use of microemulsions has been another method that is employed to form spherical nanoparticles wherein the emulsion provides a sterically enclosing environment for isolated particles to form without aggregation.

Formation of thin films and coatings: The sol-gel method has found wide applicability in the construction and formation of two-dimensional nanostructures such as thin films and coatings. The films made via the sol-gel method provide a better dimensional control with the ability to coat complex shapes, and can even be used to form multilayer coatings. In order to form a coating, a stable sol with controlled viscosity is desired. Usually stable sols can be obtained under acidic conditions and by adjusting the amount of the precursors, solvent, and pH it is possible to control the viscosity of the sol. The films can be formed

Sol-Gel Method,

Fig. 5 Schematic depiction of dip-coating (left) and spin-coating (right) processes



by dip-coating, spin-coating, or spray-coating methods. The dip-coating method of thin film deposition is perhaps more important technologically since a uniform coating can be deposited onto substrates of large dimensions and complex geometries. In the dip-coating process, a suitable substrate immersed in a low viscosity sol is withdrawn slowly at a constant speed (Fig. 5). The adhesion of the sol layer on the substrate aided by gravitational draining leads to formation of a thin film. In order to deposit uniform homogeneous films, a steady vibration-free withdrawal of the substrate from the sol is necessary. Film thickness is controlled by adjusting the viscosity of the sol and the rate of pulling the substrate. Another method of forming thin films is via spin-coating wherein an excess of sol is placed onto a substrate that is rotated at high speeds to disperse the liquid uniformly (Fig. 5). The film thickness in this case depends upon viscosity of the sol and the speed of rotation along with the rate of evaporation of the solvent. The conversion of deposited sol to a porous gel is caused mainly by solvent evaporation. Unlike the bulk gel system where gelation, aging, and drying occur sequentially over a period of several days, all of these processes typically occur within 30 s to 1 min in the thin-film such that the drying stage overlaps the gelation and aging stages. Loss of solvent molecules accelerates hydrolysis and condensation steps such that the two steps occur concurrently leading to the eventual formation of a porous xerogel structure. The important microstructural characteristics of coatings and films such as porosity, surface area, and pore size are determined by pH, concentration, and nature of precursors, viscosity of the initial sol, the solvent and its rate of evaporation, the rate of sol dispersion on the substrate, the

ambient temperature, the rate of drying, and the relative rates of evaporation and condensation reactions during drying.

Organically modified materials: Perhaps the most useful aspect of the sol-gel method is the ability to make materials with covalently integrated organic functionalities via hydrolysis of precursors containing hydrolytically stable organic groups. This enables a wide range of functional groups to be incorporated within the oxide-based network to form a diverse array of new materials with unique, novel, and interesting properties. More than one type of precursors with different functional groups can also be used to include a multiplicity of functional groups within a given material. An important consideration in use of multiple precursors is their differential reactivities toward hydrolysis and condensation which can lead to phase separation and/or spatially isolated nanodomains. Strategies to overcome these are based on prehydrolysis of the more reactive precursor followed by addition of the other precursors as well as use of different alkoxides groups on the precursors to modulate their reactivities. The organically modified materials contain less than four siloxane linkages per silicon atom and are more elastic as compared to pure oxides. The ability to tune functionality in the organically modified materials provides appealing prospects for applications based on the nature of organic group incorporated.

Hybrid Materials: Different types of hybrids can be made via the sol-gel method. These include organic-inorganic hybrids and nanocomposites. The organic-inorganic hybrids combine the properties of organic materials with the mechanical strength of the oxides. The main issue in formation of hybrids relates to their miscibility which can be enhanced by means of suitable solvents or by use of

functional groups that can noncovalently interact with silica pore structure via hydrogen bonding or electrostatic interactions. The simplest way to make a hybrid material is by encapsulation of organic molecules in the pores of the gels. The organic molecules in the pores of gels can be used to alter the properties and structure of the inorganic network. Typical examples of these include glasses containing encapsulated catalysts, dyes, and other optically active organic molecules. Another approach is based on forming interpenetrating networks of organic polymer with the silica network. These nanocomposites combine the properties of both inorganic and organic polymers, and usually influence the structure, morphology, and properties of each other. These synergistic interactions make the hybrid materials structurally and functionally superior as compared to their component parts.

An important development in formation of hybrid materials by the sol-gel method centers on formation of bio-hybrids. The solution based sol-gel pathway can be used for direct integration of biological species such as proteins, enzymes, DNA, RNA, and even whole cells. Optimized pH regions are necessary during preparation to stabilize the biological component. Typical procedure of enzyme immobilization involves initial acid hydrolysis followed by raising the pH to neutral region. The biological macromolecules are usually added to a buffered sol and get encapsulated in the growing gel network as the sol turns to a gel. The water-filled porous structure of the gel along with electrostatically charged and hydrogen bonding pore walls provide a noncovalently interacting environment for the biological molecules. The encapsulated biological entity retains its structural and functional characteristics and is able to interact with exogenous molecules that can diffuse in and out of the porous structure. The last two decades have seen tremendous advances in the design and development of bio-silica hybrids made by the sol-gel method [3].

Key Research Findings

Structure and Synthesis: The sol-gel method has been used for several decades to make silica-

based materials [6]. Oxides of different metals including mixed oxides have been prepared [7]. Initial research in this area focused on mainly pure and mixed oxides but in recent times the focus has shifted to organically modified, hybrid, and composite system. A classification of organic-inorganic materials synthesized using the sol-gel method has been developed to categorize them [8]. Nanotechnology has also spurred an active interest fabrication of sol-gel-derived particles [9]. These particles are usually synthesized in a microemulsion and can be solid, hollow, or contain core-shell structures.

Synthesis of organically modified glasses has been studied in much detail and a wide range of precursor chemistries have been identified. The organically modified precursors provide access to a diverse range of materials with tunable properties including chemical reactivity, optical properties, electronic properties, enhanced mechanical stability, and elasticity. A major area of research using these precursors has been in the construction of mesoporous materials by means of using suitable templates wherein the assembly is aided by electrostatic and hydrogen bonding interactions between the template and the growing silica network [10]. Gels made from organosilicate precursors made with spacer linkers containing both hydrophobic and hydrophilic groups exhibit hydrogel-like properties and have been shown to undergo structural and volume change with respect to different environmental stimuli. These materials are characterized by enlarged pores due the presence of spacer units along with elasticity which confers the necessary flexibility to undergo shape and volume changes [11]. These materials also provide a suitable matrix for immobilization of biological molecules.

An important aspect of protein and enzyme immobilization has been elucidation of local environment of the immobilized biomolecule and its interactions with the sol-gel matrix [12]. The immobilizing environment of the pore restricts translational diffusion, rotational mobility, and conformational flexibility. Proteins and enzymes have been shown to exhibit enhanced stability against denaturation when encapsulated in the sol-gel-derived glasses. The enzymes in the gel

matrix retain their catalytic activities and the materials have been used as solid-state catalysts for various transformations. Studies have shown that when large proteins get encapsulated in the growing gel network, they act as self-specific templates for the formation of their own pore environment. Recent research also shows that cells can be used for directing assembly of sol-gel materials. Furthermore, proteolytic enzymes have been shown to catalyze the sol-gel reactions leading to formation of gels starting from alkoxide precursors.

Function and Properties: The transparency of the sol-gel-derived glasses has been widely utilized for the development of functions related to modulation of photon flux through the materials [13]. Incorporation of dyes and luminescent fluorophores has been used as a means to develop materials for lasers, luminescent materials, and solar concentrators. Incorporation of photochromic molecules in the pores of the gels has been employed as a means to impart photochromic properties to the resulting glasses. Electronic properties have been developed in the sol-gel glasses via incorporation of conducting polymers, metal particles, or formation of mixed oxides. The nanopores of the gels have been used to form nanoscale superparamagnetic particles of transition metal oxides. The use of organically modified precursors also provides access to tailoring of mechanical properties such as strength, elasticity, and hardness depending upon the functional groups used. In addition to development of physical properties, use of precursors with specific functional groups has also been used to tune their chemical properties.

Applications: The applications of the sol-gel-derived systems have been continuously increasing [14]. The transparency of silica-based systems makes them especially suitable for development of optical applications and sol-gel-derived materials have been used as optical sensors based on changes in color or luminescence upon reaction with an analyte [15]. The analytical applications of sol-gel materials as electrochemical sensors

have also been developed [16]. Additionally, the presence of specific functional groups on organosilica materials confers selectivity in their interactions and these materials have been used as separation media in chromatography columns [17]. The use of porous particles for controlled release applications has opened up new opportunities in market and commercial products have been designed for acne treatment and sunscreen release based on particles made using the sol-gel method. A new area in this direction has been applications in fabric-care systems and environmentally sensitive organosilica hydrogels have been demonstrated as controlled release vehicles for water-triggered release of enzymes.

Coatings made from sol-gel pathways have been utilized for an array of applications including corrosion protection and surface passivation, tailoring water affinity of surfaces, enhancing optical gloss, thermal insulation, scratch resistance, and fingerprint resistance among a host of other applications [18]. The ability to modify hydrophobicity and surface microstructure of these coatings has enabled construction of surface features that mimic the lotus leaf effects. An offshoot of this research has also resulted in textile treatments to make them stain- and water-repellent.

In recent times, the use of bio-sol-gel hybrids has been an area of much activity. Initial applications related to immobilization of protein and enzymes inside the pores of silica sol-gels to make solid-state biocatalysts for various chemical transformations. The enzyme sol-gel hybrids have also been used as optical and electrochemical sensors, and as media for bio-affinity chromatography. The use of sol-gel-derived materials has also been established for immobilization of whole cells and microorganisms [19]. Organosilica gels as matrices for bio-immobilization have provided access to a wide range of biocatalysts whose activities can be exploited in practical applications. The applications domain of materials incorporating biological species continues to grow and the biological-inorganic composite area remains a topic of continued development.

Future Directions for Research

The sol-gel method of making materials provides a versatile pathway to not only making new materials, but also making known materials more useful by obtaining them in different morphologies and length scales. The last decade has witnessed significant advances in construction of nanoscale particles, assemblies, and organized structures made using the sol-gel method. Fabrication of different oxides, nanohybrids, and nanocomposites characterized by distinct nanoscale structures, surfaces, and interfaces that can be tailored to obtain the desired functional responses would continue to remain an area of vital investigation in the near future. Use of nanoscale building blocks obtained by the sol-gel method and their organization into higher-order structures along with control, manipulation, and regulation of the emergent properties at different length scales to generate specifically tailored responses would provide an appealing paradigm for future studies. An additional key issue to address in the future development of nanoscale systems made via the sol-gel method would be resolving, manipulating, and controlling the functional dichotomy of properties originating from the nanoscale as well as collective properties due to the ensemble nature of the materials.

Being able to tailor properties of materials made from molecular precursors with specific chemistries offers enticing prospects for controlling and manipulating their organization as they undergo their structural evolution from molecules to clusters to particles and other higher order structures. While there have been some advances in supramolecular assemblies of sol-gel-derived materials, the complexity and sophistication to achieve self-assembling structures with distinct interactions, interfaces, and organization over varied length scales remain as yet elusive. An area of ambitious exploration, given the formidable challenges, would be the development of new precursors and chemistries to fine-tune covalent and noncovalent

interactions to generate supramolecular systems with programmed assembly and organization. Tuning this dynamic interplay of structure and organization would require new methods and techniques of sol-gel chemistry that would provide access to far from equilibrium structures. In the near term, the structural-organizational paradigms from nature would continue to provide the necessary inspiration toward achieving these objectives.

The ability to fabricate bio-inorganic hybrids and composites has opened another exciting avenue in the design of novel materials with useful properties and applications. The integration of biological entities into the nanopores of materials, to a certain extent, restores the spatial and supramolecular organization that exists in the native cellular environment which is lacking in freely mobile biomolecules dispersed in an isotropic solution phase. The next extension of this method would generate organized assemblies that would take advantage of the specialized interactions at the biological-inorganic interfaces. Consolidated nanostructures integrating biological materials would become increasingly significant in creating a new generation of bio-hybrids that would combine the structural integrity of the sol-gel-derived materials with the functional and operational diversity of biomaterials.

During the last decade there have been significant advances in mimicking nature with sol-gel-derived materials including biocatalysis, immunodetection, surfaces that mimic lotus leaf, and biomineralized supramolecular architectures. The opportunities provided by the facile access to nanomaterials, internal porous architecture, surfaces, and interfaces would continue to evolve as chemists, materials scientists, biochemists, and physicists attempt to develop correlations between structure, morphology, form, and function. The next evolution of these materials lies in the design of nanomaterials with adaptive capabilities that can adapt to their environment and exhibit active functional responses to generate dynamically active systems capable of self-diagnosis and self-correction. The interplay of

structure, form, function, and operation that exists at different length scale would provide the guidelines for the development of next generation of materials capable of intelligent responses. While recent developments have seen the emergence of materials with smart functions, many of the systems, for the most part, remain monofunctional and passive. In the long range, design and development of new materials with hierarchical functions that span a range of functional length scales and tunable active responses would generate the next wave in this line of research.

The last decade has also seen utilization of sol-gel methods to fabricate devices for detection, diagnosis, separation, sensing, controlled release, and delivery to name a few. The real potential of these materials as nanoscale devices remains to be explored. The ability to tailor size, porosity, and the surfaces of sol-gel nanomaterials would find new avenues of research activity in the development of nanodevice technology based on a combination of inorganic, organic, and biological components. Multifunctional systems that can detect, sense, separate, sort, release, and deliver molecules to targeted sites and in a predetermined fashion would constitute the long range evolution of nanodevices made using sol-gel-derived systems. In the near term, effective utilization of strategies to adjust and modulate physical and chemical interactions of exogenous molecules with sol-gel systems would provide the necessary means of control and regulation to design nanoscale sensors, filters, separators, sorters, and delivery agents. Development of nanodevice systems for biomedical applications would also constitute an important area of research and development.

Finally, another area that would become increasingly important is the development of new “green” technologies utilizing the sol-gel method. The opportunities offered by the sol-gel method would provide the ideal framework for development of novel (bio)catalytic systems, molecular filters, sorters, or separators to reduce, remove, and/or remediate toxic and harmful substances. The sol-gel method is a greener alternative to conventional methods, however, use of volatile solvents and formation of by-products remain as issues to be

addressed. Development of new processing techniques that minimize waste and improve atom economy by eliminating by-products would be essential to the development of eco-friendly manufacturing processes based on sol-gel methods. In the long range, sustained development and refinement of the sol-gel method would be critical to displace the current manufacturing processes in large-scale production technologies.

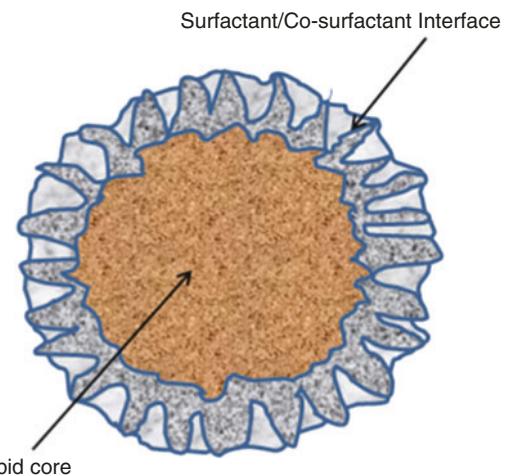
Cross-References

- [Bioinspired Synthesis of Nanomaterials](#)
- [Biosensors](#)
- [Self-Assembly of Nanostructures](#)
- [Smart Hydrogels](#)

References

1. Wright, J.D., Sommerdijk, A.J.M.: Sol-Gel Materials – Chemistry and Applications. CRC Press, Boca Raton (2001)
2. Sanchez, C., Belleville, P., Popall, M., Nicole, L.: Applications of advanced hybrid organic–inorganic nanomaterials: from laboratory to market. *Chem. Soc. Rev.* **40**, 696–753 (2011)
3. Avnir, A., Coradin, T., Lev, O., Livage, J.: Recent bio-applications of sol-gel materials. *J. Mater. Chem.* **16**, 1013–1030 (2006)
4. Hubert-Pfalzgraf, L.G.: Alkoxides as molecular precursors for oxide-based inorganic materials – opportunities for new materials. *New J. Chem.* **11**, 663–675 (1987)
5. Mutin, P.H., Vioux, A.: Nonhydrolytic processing of oxide-based materials – simple routes to control homogeneity, morphology, and nanostructure. *Chem. Mater.* **21**, 582–596 (2009)
6. Hench, L.L., West, J.K.: The sol-gel process. *Chem. Rev.* **90**, 33–72 (1990)
7. Livage, J., Henry, M., Sanchez, C.: Sol-gel chemistry of transition metal oxides. *Prog. Solid State Chem.* **18**, 259–341 (1988)
8. Sanchez, C., Ribot, F.: Design of hybrid organic–inorganic materials synthesized via sol-gel chemistry. *New J. Chem.* **18**, 1007–1047 (1994)
9. Ciriminna, R., Sciortino, M., Alonso, G., de Schrijver, A., Pagliaro, M.: From molecules to systems – sol-gel microencapsulation in silica based materials. *Chem. Rev.* **111**, 765–789 (2011)
10. Rivera-Munoz, E.M., Huirache-Acuna, R.: Sol-gel-derived SBA-16 mesoporous material. *Int. J. Mol. Sci.* **11**, 3069–3086 (2010)
11. Rao, M.S., Grey, J., Dave, B.C.: Smart glasses – molecular programming of dynamic responses in

- organosilica sol-gels. *J. Sol-Gel. Sci. Technol.* **26**, 553–560 (2003)
12. Mena, B., Mena, F., Aiolfi-Guimaraes, C., Sharts, O.: Silica-based nanoporous glasses – from bioencapsulation to protein folding studies. *Int. J. Nanotechnol.* **7**, 1–45 (2010)
 13. Penard, A.-P., Thierry, G., Boilot, J.-P.: Functionalized sol-gel coatings for optical applications. *Acc. Chem. Res.* **40**, 895–902 (2007)
 14. Aegeerter, M.A., Menning, M.: Sol-gel Technologies for Glass Producers and Users. Kluwer, Boston (2010)
 15. Tran-Thi, T.-H., Dagnelie, R., Crunaire, S., Nicole, L.: Optical chemical sensors based on hybrid organic–inorganic sol–gel. *Chem. Soc. Rev.* **40**, 62–639 (2011)
 16. Walcarius, A., Collinson, M.M.: Analytical chemistry with silica sol-gels – traditional routes to new materials for chemical analysis. *Annu. Rev. Anal. Chem.* **2**, 121–143 (2009)
 17. Kłoskowski, A., Pilarczyk, M., Chrzanowski, W., Namiesnik, J.: Sol-gel technique – a versatile tool for adsorbent preparation. *Crit. Rev. Anal. Chem.* **40**, 172–186 (2010)
 18. Aegeerter, M.A., Almeida, R., Soutar, A., Tadanaga, K., Yang, H., Watanabe, T.: Coatings made by sol–gel and chemical nanotechnology. *J. Sol-Gel. Sci. Technol.* **47**, 203–236 (2008)
 19. Livage, J., Coradin, T.: Living cells in oxide glasses. *Med. Mineral. Geochem.* **64**, 315–332 (2006)



Solid Lipid Nanoparticles (SLN), Fig. 1 Model representation of general structure of SLN

biocompatible and biodegradable lipids solid at room and body temperature.

Solid Lipid Nanocarriers

► Solid Lipid Nanoparticles (SLN)

Solid Lipid Nanoparticles (SLN)

Claudia Musicanti and Paolo Gasco
Nanovector srl, Torino, Italy

Synonyms

Lipospheres; Solid lipid nanocarriers; Solid lipid nanospheres; Solid lipid-based nanoparticles

Definition

Solid lipid nanoparticles (SLN) are drug carriers in submicron size range (50–500 nm) made of

Introduction

Solid lipid nanoparticles (SLN) were developed at the beginning of the 1990s as alternative colloidal carriers to emulsions, liposomes, and polymeric nanoparticles. SLN have attracted increasing attention as delivery system for hydrophobic drugs: prepared with lipids, they can be administered by different routes of administration such as oral, parenteral, dermal, transdermal, ocular and pulmonary.

A typical model of the structure of SLN consists of a solid lipid core surrounded by an emulsifier interface which stabilizes the particle (Fig. 1).

SLN can offer several advantages for drug delivery

1. Possibility of controlled release of drug from lipid matrix
2. Possibility of drug targeting (active and passive)
3. Protection of incorporated drug from degradation
4. Increasing drug bioavailability

5. Feasible incorporation of hydrophobic and also hydrophilic drugs (lipid drug conjugates or other techniques)
6. Very low when absent toxicity
7. Affinity for biological barriers
8. Modification of pharmacokinetic parameters and drug distribution in organs
9. Possibility of large-scale production

On the other hand SLN show certain possible limitations:

1. Poor drug loading capacity
2. Possible drug release during storage
3. Possible stability problems during long-term storage (aggregation, component degradation, tendency to form gel)

The principal components used for the preparation of SLN are:

1. Lipids, such as triglycerides (trilaurin, trimyristin, tripalmitin, tristearin), mono/di/triglycerides mixtures (glyceryl stearate, glyceryl palmitostearate, glyceryl behenate), fatty acids (stearic acid, palmitic acid, behenic acid), waxes (cetyl palmitate), cholesterol esters
2. Emulsifiers (surfactants), such as phospholipids (e.g., lecithins), polysorbates (Tween) and sorbitan esters (Span), polymers (e.g., poloxamers), generally used to stabilize interface between water and lipid
3. Co-emulsifiers (cosurfactants), such as bile salts, short-chain fatty acids and alcohols

Emulsifiers are amphiphilic molecules which possess both hydrophobic and hydrophilic portions; they are able to stabilize emulsion system by placing at the oil/water (o/w) interface (orienting the hydrophobic portion to the oil phase and the hydrophilic portion to the aqueous phase) and lowering the interfacial tension between the two phases. Co-emulsifiers intercalate between the emulsifier molecules and contribute to reduce the interfacial tension between oil and water.

Additional ingredients are used to specifically modify surface of SLN:

1. Stealth agents, such as polyethylene glycol (PEG) for improving circulation time
2. Charge modifiers to modify surface charge
3. Targeting molecules, such as peptide or antibody fragment, for linking to specific site receptor (active targeting)

Antioxidants, antimicrobial agents, and thickening agents can be usually added to SLN (dried or in dispersion) as for other pharmaceutical products.

Preparation Methods of SLN

General issue in the preparation of SLN is to obtain liquid phase where lipid and water can be emulsified before SLN formation: lipids, which are solid at room temperature, are melted or they are dissolved into solvents. Main methods to prepare SLN are reported below.

High-Pressure Homogenization (HPH)

High-pressure homogenization is a well-established technique on the large scale and it is already available in the pharmaceutical industry. High-pressure homogenizers have been used extensively in the production of nanoemulsions for parenteral nutrition.

High-pressure homogenizers push a liquid with high pressure (100–2,000 bar) through a micron size gap: by applying high pressure, the liquid accelerates to high velocity (over 1,000 km/h), and the resulting shear stress and cavitation forces break down the accelerated particles to submicron size.

HPH technique can be applied by two main different approaches: hot high-pressure homogenization (HHPH) technique and cold high-pressure homogenization (CHPH) [1].

For both techniques, the drug is dissolved or dispersed in the melted lipid at approximately 5–10 °C above its melting point.

For the HPH, the drug-loaded lipid is dispersed under high-speed stirring in a hot aqueous surfactant solution maintained at same temperature, to form a pre-emulsion. The obtained pre-emulsion is then subjected to high-pressure homogenization, still at temperature above melting point of lipid, to produce a hot o/w nanoemulsion. The homogenization process can be repeated several times: in most cases three to five homogenization cycles at 500–1,500 bar are performed. Increasing the homogenization pressure or the number of cycles often results in an increase of the particle size due to particle coalescence occurring for the high kinetic energy of the particles.

Obtained o/w nanoemulsion is then cooled down to room temperature where SLN solidify.

In CHPH, the drug-loaded lipid is rapidly cooled using liquid nitrogen or dry ice to obtain a solid solution which is then ground (milled) to obtain microparticles, approximately in the range 50–100 µm, then the solid microparticles are dispersed in a cold surfactant solution to obtain a pre-suspension. The pre-suspension is subjected to high-pressure homogenization at room temperature or below to obtain solid lipid nanoparticles: usually larger particle size and broader size distributions are obtained compared to hot homogenization.

Cold homogenization minimizes the thermal exposure of the drug, although it does not avoid it, due to dissolution of drug in melted lipid in the initial step.

Main reported disadvantages of HPH are: high energy-intensive operating conditions and potential thermal degradation of drug and excipients during production process.

Warm Microemulsion Technique

Microemulsions are transparent (clear), optically isotropic and thermodynamically stable systems. They are dispersions of two immiscible liquids (oil and water) stabilized by surfactants and optionally by cosurfactants.

Depending on the composition of the system different structures can be formed:

1. Oil dispersion in water medium – o/w microemulsion
2. Water dispersion in oil medium – w/o microemulsion
3. Bicontinuous structure

In o/w and w/o microemulsions, the dispersed phases are in shape of very small drops (10–100 nm) stabilized by an interfacial film of surfactant and cosurfactant molecules.

In bicontinuous structures, regions of water and oil are interdispersed with no spherical geometry: it occurs when the amount of water and oil are comparable.

The thermodynamic stability of microemulsion is due to the very low interfacial tension which allows the spontaneous formation of microemulsions without high energy input.

Warm microemulsions obtained with melted lipids have been used as precursor for the preparation of SLN: after their preparation at temperature ranging from 55 °C to 85 °C depending on melted lipid used, they are then dispersed in cold aqueous medium where melted lipid drops solidify into SLN [2].

The hydrophobic drug is dissolved in mixture composed of melted lipid and surfactant maintained at warm temperature, the cosurfactant is dissolved in water and the obtained solution is heated at the same temperature of the lipid mixture; cosurfactant aqueous solution is then added to the mixture composed of lipid, drug and surfactant and by mild mixing an o/w microemulsion is obtained and successively dispersed in cold water at 2–3 °C under mechanical stirring.

Also hydrophilic drugs have been loaded into SLN by water in oil in water (w/o/w) double microemulsion technique: water, where hydrophilic drug is dissolved, constitutes the internal phase of w/o microemulsion, and the melted lipid (55–85 °C) the external phase; the w/o warm microemulsion is then added to an external aqueous solution of surfactant and cosurfactant maintained at the same temperature and after mixing w/o/w microemulsion is obtained and dispersed in cold water to obtain SLN.

The SLN dispersion, obtained by warm microemulsion technique, is usually purified by tangential ultrafiltration in order to remove surfactant and cosurfactants not incorporated into SLN, and to concentrate dispersion when needed.

Advantage of the production of SLN by warm microemulsion include: low mechanical energy input, flexibility of interphase composition that allow surface functionalization, regular spherical shape. Potential drawbacks are thermal exposure, which can cause degradation of drug and of excipients, use of higher amount of surfactant and cosurfactant, which sometimes need to be removed.

High Shear Homogenization and Sonication

Following this process, drug dispersed in melted lipid phase is homogenized by rotor-stator homogenizer with a hot surfactant water solution to obtain nanoemulsion that is then cooled to form SLN.

Process parameters that can affect particle size are: emulsification time, stirring rate, and cooling conditions.

Homogenization can be used in association with sonication: melted lipid phase loaded with drug is homogenized with a hot surfactant solution to obtain a coarse emulsion that is then ultrasonicated to obtain final nanoemulsion before SLN are formed upon its cooling to room temperature [3].

Main possible disadvantages of this method include high energy input, broad particle size distribution, and potential damage of sensitive biomolecules.

When sonication is used, potential metal contamination and temperature increase have to be considered.

Solvent Emulsification-Evaporation Method

In this method, drug, lipid matrix, and emulsifier are dissolved in a water-immiscible organic solvent which is then emulsified in an aqueous phase containing a water-soluble cosurfactant and homogenized. Upon evaporation of the solvent under reduced pressure, SLN are formed by precipitation of the lipid in the aqueous medium [4].

Parameters that affect mean particle size of nanoparticles obtained by this method are: surfactant/cosurfactant blend and lipid concentration in the organic phase.

Advantages of this technique are: avoidance of heat application, production of small size nanoparticles that can be sterilized by filtration through 0.2 µm pores.

The major possible disadvantage of this method is the presence of organic solvent residues which can cause toxicity.

Solvent Emulsification-Diffusion Method

The process consists in dissolving the lipid in a partially water-soluble solvent (previously saturated with water) at room or controlled temperature, depending on solubility of lipid in the solvent. This organic phase is emulsified with an aqueous solution (saturated with solvent) containing the stabilizing agent and maintained at the same temperature, by a rotor-stator homogenizer. This o/w emulsion is then diluted with water maintaining a constant stirring and controlled temperature for promoting the diffusion of solvent of the internal phase toward the external phase, causing lipid aggregation and precipitation in form of SLN.

Depending on its boiling point, solvent can be eliminated from SLN dispersion under reduced pressure or by washing (ultrafiltration).

The selection of the water-miscible solvent and the stabilizers are critical parameters to obtain lipid particles in the nanometric range: usually solvents with high water miscibility and stabilizers able to form stable emulsions are preferred.

It is possible to reduce the particle size by increasing the process temperature, the stirring rate, the amount of stabilizer, and by lowering the amount of lipids [5].

This technique is efficient and easy to be implemented, so feasible for industrial upscaling: no high energy is required and low physical (thermal and mechanical) stress is applied.

Possible drawbacks of this method are solvent residues which need to be cleaned up and large dilution produced to obtain diffusion of solvent which needs further concentration steps.

Solvent Injection Method

In this procedure, solid lipid is dissolved in a water-miscible solvent or solvents mixture, and then rapidly injected through an injection needle into a stirred aqueous phase with or without surfactants.

The SLN production by this technique is based on the rapid diffusion of the solvent across the solvent-lipid and solvent-water interface, causing precipitation of nanoparticles.

Two simultaneous effects contribute to the effective formation of SLN:

1. Gradual solvent diffusion out of lipid-solvent droplets into water causes reduction of droplet size and simultaneously increases lipid concentration
2. Diffusion of pure solvent from the lipid-solvent droplet causes local variations in the interfacial tension at droplet surface, inducing reduction of size of droplets

In this process particle size of SLN can be influenced and controlled by variation of process parameters such as injected solvent, lipid concentration, injected volume of solvent, lipid concentration in the solvent phase and viscosity of the aqueous phase [6].

This technique is simple and fast, without the need of sophisticated equipment; possible disadvantage of this method is the use of organic solvent which has to be considered for pharmaceutical applications and for this reason a complete removal of the solvents by ultrafiltration, evaporation, or freeze-drying is required.

Coacervation Method

This process allows to obtain SLN made of fatty acids by acidification of micellar solution of their alkaline salts: fatty acids sodium salts are dispersed in a polymer water solution (use of steric stabilizer polymer is essential to avoid particle aggregation) and heated under stirring just above the Krafft point of the fatty acid sodium salt to obtain a clear solution (Krafft point is the temperature at which the solubility of sodium soap increases dramatically and the solution becomes isotropic and transparent): an acidifying solution

is then added drop-wise until pH about 4 is reached and the obtained suspension is then cooled down to 15 °C in a water bath under stirring.

Parameters that can influence size of nanoparticles and polydispersity index are: type and molecular weight of polymer and lipid concentration in the dispersion [7].

Although average dimensions are reported in a bigger range of 260–500 nm, this process presents several advantages like absence of solvents and need for common apparatus: the method is feasible and suitable for laboratory production and it is easy to be scaled up; studies are ongoing for loading of different drugs.

Membrane Contactor Method

This method employs a cylindrical membrane module: aqueous phase containing a surfactant is circulated in internal channel of membrane and melted lipid is pressed through pores of membrane into internal water flow allowing the formation of small droplets which are swept away by the aqueous phase; water is maintained at lipid melting temperature. SLN are then formed by cooling the preparation to room temperature.

The membrane contactor allows the preparation of SLN with a lipid phase flux between 0.15 and 0.35 m³/h/m² [8].

The advantages of this new process are its simple use, the control of the SLN size by an appropriate choice of process and membrane parameters, and its scaling-up abilities. Drug loading and surface modification, for targeting needs, have to be fully assessed and developed.

S

Characterization of SLN

Particle Size, Size Distribution and Shape

Particle size, shape and stability over time may influence several important pharmaceutical features of SLN such as drug release, suitable administration route (e.g., particle size is a critical issue for intravenous administration, where particle size must not aggregate or enlarge by protein binding to exceed diameter of blood vessels) and in vivo biodistribution.

Particle size and morphology of SLN are mainly affected by method of production and composition (lipid matrix, type and amount of emulsifiers, viscosity of aqueous phase, etc.).

Main techniques commonly used for characterization of SLN are summarized below.

Photon correlation spectroscopy (PCS) also known as dynamic light scattering (DLS), is widely used method for the measurement of particle size in range from nanometer to few micrometers.

This technique determines hydrodynamic diameter of particles by measuring fluctuation of the intensity of scattered light which is due to Brownian motion of particles.

Instruments usually report hydrodynamic diameter (Zaverage) and polydispersity index (PI), which give estimation of width of particle size distribution.

Electron microscopy methods are very useful techniques for determining shape, morphology and size of lipid nanoparticles.

The most commonly reported method for morphological characterization is *transmission electron microscopy* (TEM): different techniques can be applied for sample preparation (negative staining, freeze-fracture, sample vitrification (cryo-TEM)) and can provide different information about colloidal lipid structures; sample preparation is a crucial point as it can lead to structural alterations of the sample that need to be taken into consideration. As an example, TEM micrographs of SLN (obtained by warm microemulsion method) are reported in Fig. 2 [9]: images show SLN in water dispersion before administration and same SLN in biological fluids coming from *in vivo* sampling after administration.

Other methods to determine SLN particle size and morphology are based on *scanning electron microscopy* (SEM).

Nonconductive samples, such as lipid-based nanoparticles, may be visualized without metal coating using specialized SEM instrumentation such as *environmental SEM (ESEM)* in which the sample is placed in a internal chamber at higher pressure than vacuum: positively charged ions generated by beam interactions

with the gas, present in the chamber, help to neutralize the negative charge on the surface of sample.

In *field emission SEM (FESEM)*, the electron beam is produced with a cold cathode field emitter instead of a thermoionic emitter (tungsten filament heated with an electric current) as in conventional SEM: this allows for much higher resolution showing particle shape and size. As an example, images obtained from two different formulations of SLN, produced by warm microemulsion technique, are reported in Fig. 3.

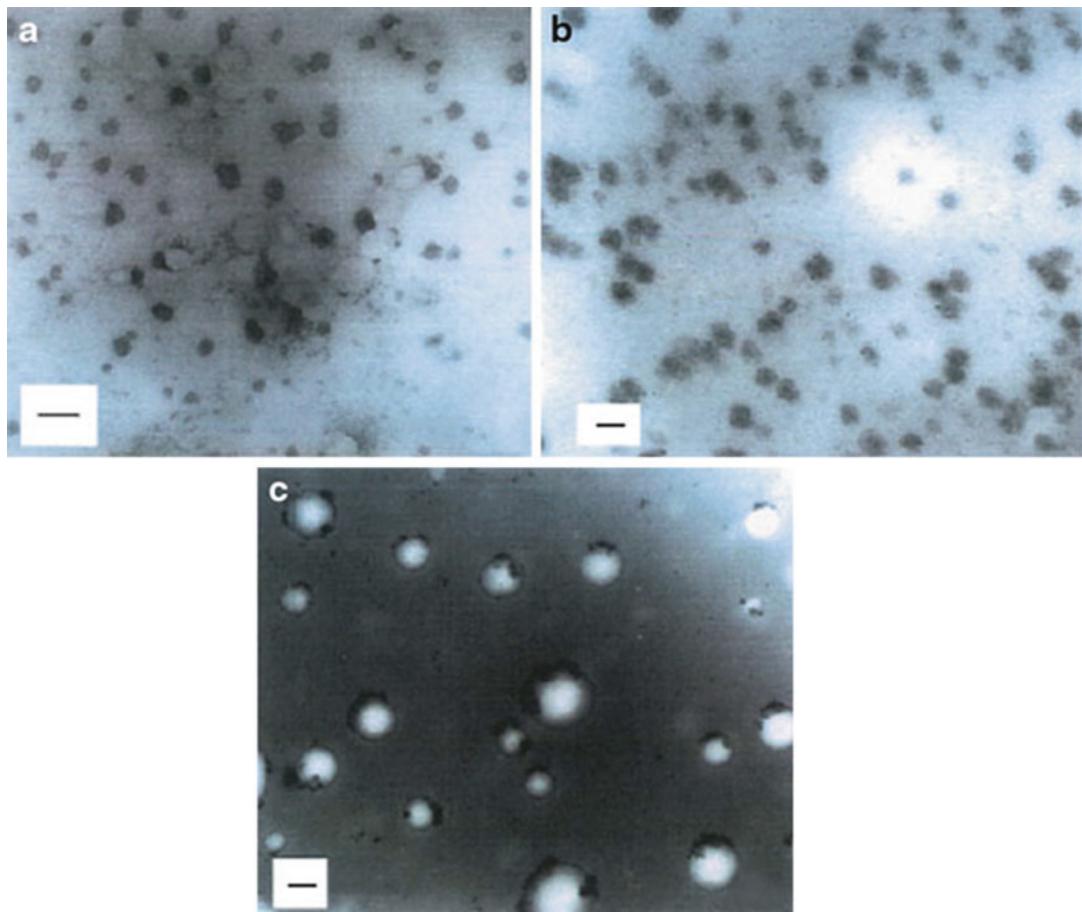
Atomic force microscopy (AFM) is another technique used to determine particle size and shape. AFM can be operated in different modalities depending on the specific application requirements: sample does not require any special treatments (such as coating) and technique also allows to work in ambient conditions, even in liquid environment. As an example, AFM characterization of SLN produced by solvent-diffusion method is reported in Fig. 4 [10]: SLN produced in miniemulsion used as nanoreactor system is compared with SLN obtained by conventional method.

Surface Charge

Nanoparticles dispersed in a liquid medium are surrounded by an electrical double layer composed by an inner layer (stern layer) where the ions are strongly bound to particle surface and an outer layer (diffuse layer) where ions are less firmly associated with the particle.

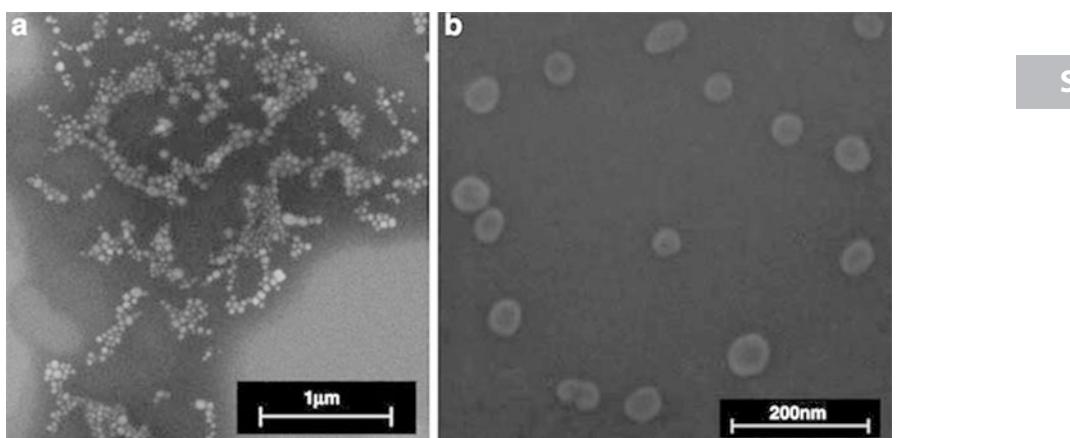
Within the diffuse layer there is a boundary (shear plane) inside of the ions and the particle form a stable entity: when the particle moves, only ions within the boundary move with the particle, while ions beyond the boundary remain in the bulk dispersant.

Zeta potential (Z potential) is the electric potential at this boundary and it gives measure of surface charge. It is used to assess stability of colloidal systems as it can express tendency of particles to repulse or to attract: higher values of zeta potential (absolute value > 30 mV) usually imply more stable dispersions due to electrical

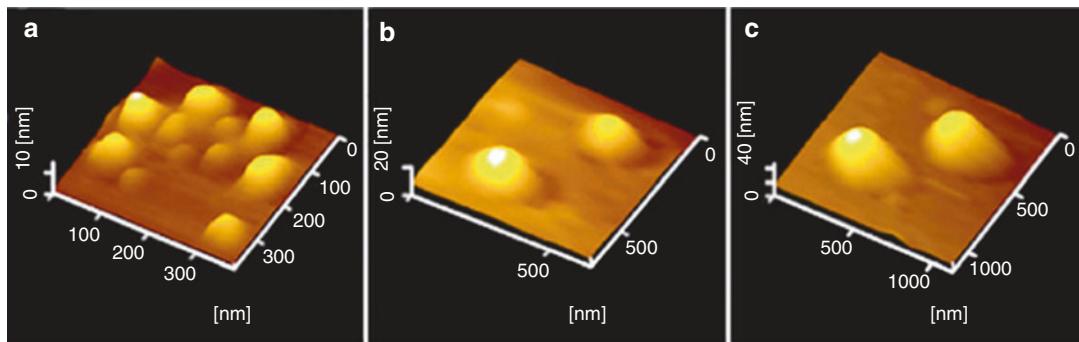


Solid Lipid Nanoparticles (SLN), Fig. 2 TEM micrographs of SLN obtained by warm microemulsion method and containing 2.5 % of tobramycin in different

physiological media: (a) aqueous dispersion before administration to rats, (b) lymph, and (c) plasma after duodenal administration to rats (bar = 100 nm) [9]



Solid Lipid Nanoparticles (SLN), Fig. 3 FESEM image of SLN obtained by warm microemulsion method: (a) SLN with lipid matrix of Stearic acid, (b) SLN with lipid matrix of cholesterol butyrate



Solid Lipid Nanoparticles (SLN), Fig. 4 Atomic force microscopy images of SLN prepared by solvent diffusion method in a nanoreactor system. (a) Blank SLN, (b) SLN

with 10 % of clobetasol propionate charged (c) SLN prepared by conventional method with 10 % of clobetasol propionate charged [10]

repulsion, while low values can indicate possible colloid instability which could lead to aggregation.

Z potential affects not only dispersion stability but also interaction with the biological surroundings (interaction with proteins and cells) and the *in vivo* fate of nanoparticles (biodistribution); it depends on the composition of the particle and on its surrounding medium: pH, electrolyte concentration, and concentration of components of the formulation.

Z potential is determined by measuring the velocity of the particles dispersed in a liquid medium under the influence of an applied electric field (electrophoresis measurements): charged particles move toward the electrode under applied electric field with velocity dependent on the strength of electric field or voltage gradient, the dielectric constant and the viscosity of the medium, and the Z potential of the particle.

Electrophoretic mobility, defined as the velocity of the particle in a unit electric field, is related to Z potential by Henry equation and is usually measured by light scattering technique.

Crystallinity, Polymorphism, Structure, and Stability

Crystallinity and polymorphic transitions of lipids in the dispersed state may differ from

bulk material due to the small particle size of colloidal system, the presence of emulsifiers used for the stabilization of SLN and the preparation method. In thermal analysis SLN dispersions usually show lipid melting point slightly shifted to lower temperatures and broader melting peak than the bulk lipid, polymorphic transitions are generally faster in SLN than in the bulk material [1, 11].

Crystallization and polymorphic behavior of SLN are correlated with drug incorporation and drug release: highly ordered crystal lattices (monoacid triglycerides) in SLN matrix can lead to drug expulsion, while less ordered crystal lattice (mixture of mono-, di-, and triglycerides), because imperfections which provide space to accommodate the drug, can allow better drug incorporation.

Polymorphic transitions may cause alterations of lipid packing and thus of the internal structure of the nanoparticles, which may have negative consequences for drug loading. The course of polymorphic transitions depends on the type of lipid matrix and can be modified by other components of the dispersions [11].

As an example, triglycerides crystallize in three main polymorphic forms: α , β' , and β ; the β form is highly ordered with high thermodynamic stability, the α form is a less ordered form with low thermodynamic stability, the β' form has

intermediate characteristics between α and β forms. SLN based on triglyceride matrix tend to crystallize in the metastable α form: generally thermodynamically unstable configurations allow lipid molecules to have higher mobility causing lower density and a higher capability to incorporate drug molecules, but during storage the α form transform via β' form in the more thermodynamically stable configuration β , characterized by higher packing density which can cause expulsion of the drug form structure.

SLN may change their shape during polymorphic transitions: as an example, tripalmitin SLN have a spherical shape in α form, but they have a platelet shape in the β form. During storage the polymorphic transition from the α to the β configuration is connected with an increase of the particle surface due to preferred formation of platelets; surfactant molecules could not provide any more complete coverage of the lipid surface and particle aggregation can occur, leading to gelation (irreversible transformation of SLN dispersion into a viscous gel) [1].

SLN can display a lower crystallization tendency than the bulk material and may not crystallize properly after preparation and storage, at a temperature below melting point of lipid, forming super-cooled melts, which are liquid lipid nanoemulsions and not solid lipid dispersion. This phenomenon is particularly pronounced in SLN made from short-chain monoacid triglycerides such as trimyristin and trilaurin, and it is mainly due to size dependency of crystallization process that requires a critical number of nuclei to start, which cannot be reached in small droplets: tendency to formation of super-cooled melts increases with decreasing of droplet size [1, 11].

Hence, the physical state is a very important parameter which affects performance of SLN as drug delivery system: the techniques usually used to investigate physical state of SLN are *differential scanning calorimetry* (DSC) and *X-ray diffraction* (XRD) techniques.

DSC is a thermal analysis technique which measures the difference in heat flow between the

sample and a reference when they are both subjected to the same controlled temperature program: DSC quantifies the enthalpy changes during endothermic and exothermic process and it is a useful technique to determine melting point, melting enthalpy, degree of crystallinity, glass transitions of amorphous material, and to identify different polymorphic forms.

In XRD techniques, an X-ray beam is focused on the sample, the scattered X-rays interfere with each other giving particular diffraction patterns: crystalline materials display many diffractions bands which depend on the disposition of the atoms within the unit cell of the crystalline lattice, while amorphous compounds present more or less a regular baseline.

XRD techniques allow to differentiate between crystalline and amorphous material and to identify different polymorphic forms in SLN dispersions: the most commonly used are *powder X-ray diffraction* (PXRD), *small angle X-ray scattering* (SAXS), and *wide angle X-ray scattering* (WAXS).

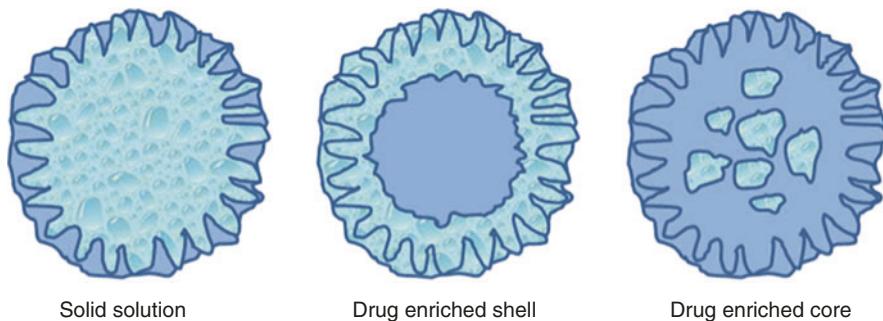
Nuclear magnetic resonance (NMR) is another useful technique to investigate structure of SLN for surfactant distribution on particle surface, drug distribution in the particle and potential formation of super-cooled melts.

Due to different chemical shifts it is possible to attribute the NMR signals to particular molecules or their segments (functional groups); NMR active nuclei of interest are ^1H , ^{13}C , ^{19}F , and ^{31}P . Width and amplitude of NMR signals can be related to liquid and semisolid/solid state of nuclei: nuclei in the liquid state give sharp signals with high amplitude, while nuclei in the semisolid/solid state give weak and broad signals [1].

S

Drug Incorporation

A broad range of drugs, mainly with hydrophobic properties, has been already incorporated into SLN, although the crystalline nature of the lipid matrix can offer limited space for drug incorporation inside the particle core [11].



Solid Lipid Nanoparticles (SLN), Fig. 5 Schematic representation of drug incorporation models for SLN

Drug loading capacity of SLN is defined as the percentage of drug incorporated into nanoparticles referred to the total weight of the lipid phase or to the content of dispersed material; value of this parameter is usually in range 1–20 %.

Drug entrapment efficiency is defined as the percentage of drug incorporated into nanoparticles referred to the total drug added for the SLN preparation; values of this parameter is usually relatively high (80–99 %).

The distribution (localization) of the drug within SLN structure (Fig. 5) can vary considerably and three *drug incorporation* models have been proposed [12]:

1. *Solid solution model.* The drug is molecularly dispersed in the solid lipid matrix (SLN matrix as a solid solution).
2. *Drug-enriched shell model.* The drug is concentrated in the outer shell of the SLN. This situation can be explained by a higher solubility of the drug in the aqueous-surfactant outer phase at increased temperature in the production process; during cooling, lipid in the core starts to solidify and becomes less/not accessible for the repartition of the drug into the lipid inner phase leading to enrichment in the particle shell.

High temperatures employed in the preparation process can increase solubility of the drug in the aqueous-surfactant phase promoting drug localization at the surface region.

3. *Drug-enriched core model.* The drug is concentrated in the core of the lipid particle and it is surrounded by a lipid shell. This situation can be explained by a precipitation of the drug before the lipid crystallizes during cooling: this takes place preferentially when drug concentration in the lipid at process temperature is at its saturation solubility, so during cooling a super saturation and subsequent drug precipitation are achieved

Drug-loading capacity of SLN is affected by different factors:

1. Solubility of the drug in the melted lipid. High solubility is a prerequisite to obtain sufficient drug loading; the amount of drug that can be dissolved in the lipid formulation may exceed the pure value of its solubility in lipid alone and a greater solubility in the whole formulation suggests that the drug can localize both in inner lipid phase and in external part on SLN (interfacial region) where surfactants are present. Solubility of drugs in the lipid can be improved by preparing lipid prodrugs such as stearic or palmitic acid derivatives (lipid drug conjugates).
2. Physical structure of solid lipid matrix. Complex lipids which form less ordered crystals with many imperfections can favor drug incorporation.

3. Polymorphic state of lipid matrix. Lipid transformation in more stable form reduces the number of imperfections in crystal lattice and can determine drug expulsion.

Determination of drug-loading capacity can be usually performed by separation of free drug from the dispersion medium by employing different techniques such as ultrafiltration (tangential or centrifugal), ultracentrifugation, gel filtration chromatography, and dialysis.

Drug Release

Drugs incorporated into SLN are released by degradation and surface erosion of the lipid matrix and by diffusion of drug molecules through the lipid matrix. SLN are composed of physiological lipids for which metabolic pathways exist in the body: most important enzymes involved in *in vivo* SLN degradation are lipases, which are present in various organs and tissues [1].

Drug release from SLN is mainly affected by localization of the drug [12]:

1. Localization of the drug within the core of solid lipid matrix offers the possibility to obtain a prolonged drug release.
2. Localization of drug molecules on particle surface often leads to burst effect (fast initial drug release). SLN can show a biphasic drug release profile: an initial burst release, due to the drug localized at the surface, is followed by a more gradual release due to the drug localized in the lipid matrix.

Possible presence of alternative colloidal species always has to be taken into account for drug release characterization: stabilizing agent cannot be localized exclusively on the lipid surface as expected, but also in the aqueous phase forming micelles, mixed micelles or liposomes that can solubilize the drugs and constitute alternative drug incorporation sites [1].

Composition of SLN is other important parameter that can affect drug release: *in vitro* experiments indicate that SLN show different

degradation rates by lipases as a function of their composition (kind of lipid matrix and emulsifier). Degradation of triglycerides SLN show dependence to the length of the fatty acid chain: longer fatty acid chain shows slower degradation, while degradation of SLN made of waxes (cetylpalmitate) is slower compared to glyceride matrices [1].

Emulsifiers containing polyethylene glycol (PEG) chains can reduce SLN degradation by a hindering effect which reduces the anchoring of the enzyme on the SLN surface [1].

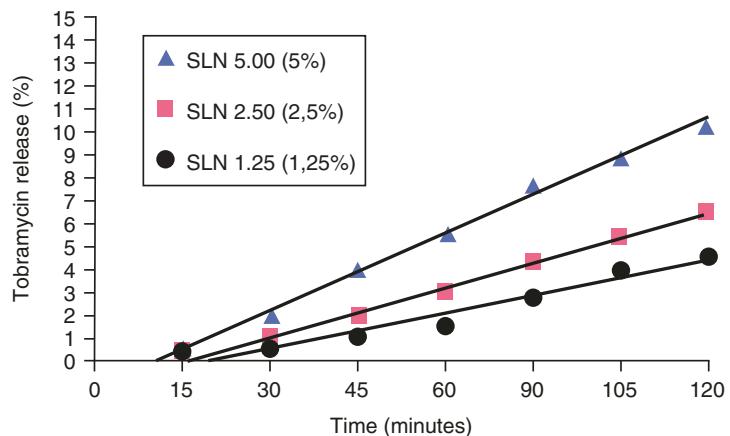
Since SLN are degraded by surface erosion, in SLN with small size and higher surface area, drug release is expected to be more rapid; nonspherical particles, such as thin platelet shape, are characterized by larger surface area and require shorter time for degradation and drug diffusion to particle surface [11].

Drug release can be studied by different techniques such as dialysis membranes (flat or bag) and Franz diffusion cells. Release kinetics is affected by *in vitro* release conditions: sink or non-sink conditions, release medium, dilution of nanoparticles dispersion that has to be considered in particular if the formulation is intended for oral and intravenous administration; release experiments have to be performed mimicking *in vivo* conditions in order to avoid distorted release profile [11].

As an example of *in vitro* study [9], drug release was investigated, working in sink conditions, for three types of SLN incorporating different percentage of tobramycin (1.25 %, 2.50 %, and 5.00 %) by using a multicells rotating tool, where donor and acceptor compartments were separated by double hydrophilic/hydrophobic membranes: results showed that release kinetic was pseudo-zero order for all three types of SLN and the amount of drug released was higher for Tobra-SLN 5.00 % than for Tobra-SLN 2.50 % which was higher than for Tobra-SLN 1.25 %, showing that *in vitro* release can be modified by changing amount of drug loaded into SLN and consequently their physical characteristics (Fig. 6) [9].

Solid Lipid Nanoparticles (SLN)

(SLN), Fig. 6 In vitro release – percentage of tobramycin diffused through a double membrane versus time for SLN loaded at three different percentages of tobramycin [9]



Routes of Administration/Applications

SLN have been tested for drug delivery applications by different routes of administration including intravenous, oral, ocular, topical dermal, transdermal, and pulmonary.

Parenteral Administration: Intravenous Route (IV)

Overview

The evaluation of SLN interaction with blood and other tissues is very important for IV administration: many studies have been performed to assess toxicity of SLN both in *in vitro* and *in vivo* tests.

Cellular toxicity of SLN has been investigated using several cell lines: *in vitro* experiments demonstrate that toxicity is dependent on composition of SLN (nature of lipid matrix and type of surfactant used) and on concentration of SLN in the culture medium [13].

In vivo toxicity studies confirmed that SLN are well tolerated in living system after their iv administration [13].

In order to obtain SLN suitable for parenteral administration, pharmaceutically acceptable excipients must be employed, and sterility must be assessed. Studies showed some SLN formulations can be sterilized by sterilizing filtration (such filtration is applicable for particles with size <0.2 µm), some other can be sterilized by

autoclaving generally at 121 °C for at least 15 min (SLN melt during autoclaving and recrystallize during cooling); physicochemical properties of SLN formulation should not change during the sterilization process. Autoclaving temperature can promote chemical degradation and can affect physical stability of SLN because their structure can be lost when particles melt and recrystallize in noncontrolled way; performed studies show that critical parameters for SLN sterilization by autoclaving are temperature and timing of sterilizing process and composition of formulation [1].

Generally, all *in vivo* studies performed showed drug-loaded SLN change pharmacokinetic parameter of carried drug when compared to reference drug solution: much higher mean residence time (MRT) and area under the curve (AUC) induced by SLN formulation mean higher drug availability and consequent possible higher efficacy.

Performed studies confirmed as well that important limiting factor for IV administration of SLN is the uptake of particles by macrophages of reticuloendothelial system (RES). SLN are recognized as foreign substances and quickly removed from blood circulation. Colloidal particles interact with blood plasma proteins called opsonins which are adsorbed on particle surface (opsonization): opsonins mediate RES recognition by interacting with specific membrane receptors of macrophages. The capacity of nanoparticles to avoid opsonization and

consequent macrophages uptake depends on size of particles, on their surface charge, and hydrophobic characteristics.

In order to avoid RES recognition, SLN surface can be functionalized with different molecules having hydrophilic and flexible chains which form a hydrophilic steric barrier that cover and protects nanoparticles from interaction with plasma proteins and prolongs their blood circulation time. Such sterically stabilized nanoparticles (now reported as *stealth*) are mainly obtained by using lipid derivatives of polyethylene glycol (PEG): amount of stealth agents and length of their chain can affect degree of surface coverage and rate of uptake by macrophages.

Macromolecular drugs and colloidal drug carriers such as SLN, liposomes, and polymeric nanoparticles, can accumulate preferentially in tumor tissue, a phenomenon called enhanced permeation and retention effect (EPR effect), due to the presence of a discontinuous endothelium and to the lack of efficient lymphatic drainage in the tumor tissue: by reducing RES uptake, nanoparticles can further passively accumulate inside such tumor tissues (*passive targeting*).

In order to further increase drug accumulation in target tissue or to produce higher and more selective therapeutic activity, specific functionalization of colloidal drug carriers has been developed (*active targeting*) by decorating SLN surface with ligand molecules able to recognize specific site on target cells inducing internalization of carrier after binding. Targeting molecules include: monoclonal antibodies or fragments, peptides, glycoproteins, and receptor ligands.

Main Applications of SLN by Parenteral/IV Administration

Cancer Therapy

SLN have been investigated as drug delivery system by IV administration mainly for cancer treatment: several antitumor drugs have been incorporated into SLN such as doxorubicin, camptothecin, paclitaxel, etoposide, idarubicin, vinorelbine, mitoxantrone, 5-fluorouracil, and

their in vitro and in vivo distribution and efficacy have been evaluated [13].

SLN are considered suitable potential carriers in oncology for ability to incorporate antitumor drugs with different chemical properties for improvement of drug stability (drug protection), enhancement of drug efficacy, reduction of drug toxicity (lowering of side effects), improvement of pharmacokinetics parameters, passive targeting, and the possibility to reach difficult districts like the brain [2, 13].

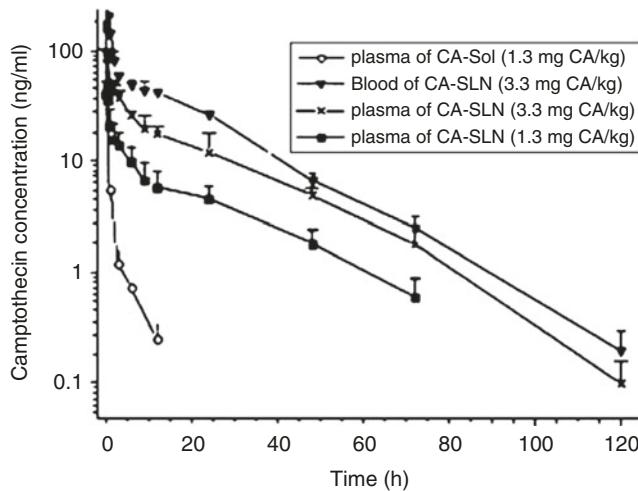
One of the first in vivo studies on these carriers in oncology has been performed on camptothecin (CA) loaded into SLN prepared by high pressure homogenization method and administered intravenously into mice: the concentration of camptothecin was determined in various organs and compared to a camptothecin control solution [13]. The results showed that in tested organs, the area under the curve (AUC) versus dose and the mean residence times (MRT), pharmacokinetic parameters relevant to drug bioavailability, of CA-SLN were much higher than those of CA-Solution (Fig. 7) [13].

As another example, in vivo tests have investigated pharmacokinetics and tissue distribution of doxorubicin incorporated in both stealth and not stealth SLN prepared by warm microemulsion technique and injected intravenously into rabbits [2]. AUC increased as a function of the amount of stealth agent present in SLN and doxorubicin was still present in the blood 6 h after the injection of both stealth and non-stealth SLN, while none was any more detectable after IV injection of doxorubicin solution.

The pharmacokinetics and tissue distribution of doxorubicin SLN were also studied in rat animal model after intravenous administration and compared with commercial solution of doxorubicin [2]: doxorubicin was still present in the blood 24 h after the injection of stealth and non-stealth SLN, while it was not detectable after the injection of the commercial solution. The results confirmed the prolonged circulation time of SLN, in particular SLN decreased the concentration of doxorubicin in the heart indicating lower cardiotoxicity compared to doxorubicin solution.

Solid Lipid Nanoparticles (SLN),

Fig. 7 Concentration-time curves of camptothecin after IV administration of CA-SLN with doses of 1.3 (■) and 3.3 (x) mg CA/kg in plasma and 3.3 (▼) mg CA/kg in blood, and CA-SOL with a dose of 1.3 (○) mg CA/kg in plasma. Results represent means \pm SD of four animals [13]



Diseases of the Central Nervous System (CNS)

Pharmacological treatment of diseases of the CNS, such as brain tumor, neurological, and neurodegenerative diseases, is limited by the presence of the *blood-brain barrier* (BBB) that restricts enormously the transport of many important drugs from the blood into the brain. The BBB is formed by the endothelial cells of the cerebral capillaries which differ from endothelial cells in the rest of the body for the absence of fenestrations, for the presence of particular and more extensive tight junctions, and for minimal pinocytic vesicular transport.

In order to evaluate transport of SLN across the BBB, both stealth and non-stealth radiolabeled SLN, prepared by warm microemulsion method, were injected intravenously into rats [2], and tissue distribution was monitored for 60 min: radioactivity in liver and in lung was lower for stealth SLN than for non-stealth SLN confirming difference in their uptake. Both types of SLN were detected in brain and cerebrospinal fluid, although at low percentage.

SLN incorporating baclofen have been investigated as new pharmaceutical preparation of this drug which is used in treatment of spasticity [2]: SLN were prepared by warm microemulsion method and were injected intraperitoneally to rats at increasing dosage. As an important finding, drug effects were detectable with lower doses of baclofen when loaded into SLN, in comparison

with the needed amount of baclofen when in solution.

In vivo results showed a good correlation with plasma and tissue concentration of baclofen: after 2 and 4 h, only baclofen-SLN produced detectable baclofen plasma concentrations, while 2 h after the administration of baclofen solution, the amount of baclofen in plasma was undetectable. Moreover, baclofen concentration in the brain 2 h after SLN administration was almost double than after baclofen solution, suggesting that baclofen may pass the BBB in much higher amount when formulated in SLN.

In order to improve brain uptake, SLN modified with specific ligands on surface have also been tested (active targeting): the presence in the BBB of receptor-mediated transport systems for endogenous molecules can be exploited to gain access to the brain.

Transferrin-conjugated SLN have been investigated for their ability to target quinine hydrochloride to brain [13]: biodistribution studies showed that quinine hydrochloride concentration in brain was significantly higher in case of transferrin-conjugated SLN as compared to that of quinine-SLN not functionalized and of quinine hydrochloride solution.

Imaging

Imaging technologies are important tools for diagnosis of diseases and for monitoring of therapies.

SLN have been proposed as carriers for hydrophobic imaging agents in order to increase concentration of the agent in the target tissue.

SLN have been investigated as carrier of contrast agents for magnetic resonance imaging (MRI) such as iron oxides nanoparticles (NP) [2]. SLN loaded with iron oxides NP were prepared by warm microemulsion method and were studied in *in vitro* and *in vivo* tests and compared to Endorem®, commercial preparation of iron oxides NP. Iron oxide-SLN showed *in vitro* relaxometric properties similar to those of Endorem® and after intravenous administration to rats showed to have slower blood clearance than Endorem®.

Gene Therapy

A potential approach for the treatment of human genetic disorders is gene therapy. This is a technique whereby the absent or faulty gene is replaced by a working gene, so that the body can make the correct enzyme or protein and consequently eliminate the root cause of the disease. Process of introducing nucleic acids into cells is generally called transfection.

Polynucleotides molecules (DNA or RNA) are large, hydrophilic, and negatively charged molecules. They are very labile in biological environment and their spontaneous entry inside cells is a very inefficient process.

Nucleic acids can be delivered by two main classes of vectors: viral vectors and nonviral vectors [14]. Although viral vectors are very efficient in nucleic acid delivery, their immunological risk is high (insertional mutagenesis, adverse immunogenic responses, inflammatory reactions) and there is limitation on dimension of nucleic acid to be incorporated: for these reasons nonviral carriers can offer preferable alternative. Most important tested nonviral vectors are liposomes, complexes made by DNA or RNA and cationic polymers or cationic lipids, and SLN [14].

The first obstacle for systemic delivery of DNA or RNA molecules is the extracellular environment, extreme pH, proteases and endonucleases, and immune defense.

The second barrier is cellular membrane; nucleic acids alone are unable to cross it because of their high negative charge.

One of the main routes of internalization of vectors carrying RNA or DNA is endocytosis: the internalized particle exists in endosomes, compartments of the endocytic membrane transport pathway, that either fuse with lysosomes, the main hydrolytic compartment of the cell where RNA or DNA are degraded losing their activity. Therefore, escape from endosomes is crucial for efficient transfection. Certain lipids have the ability to destabilize endosomal membranes favoring the escape of nucleic acids from this compartment: those lipids are called fusogenic lipids (e.g., dioleoyl phosphatidylethanolamine DOPE) and they are usually included in SLN formulation for this reason.

For their interference with lysosomes, lysosomotropic agents, such as cloroquine, are reported as well to enhance gene expression [14].

Another barrier against DNA transfection is nuclear envelope, containing pore regulating passive transport for molecules with maximum mass of 70 kDa or mean diameter size of 10 nm.

Cationic SLN, with positive surface charge, carry the nucleic acids by means of electrostatic interaction: different *in vitro* studies showed their efficacious transport of nucleic acid through cellular barriers into the cytoplasm or nucleus [12, 14].

Cationic SLN can be obtained by using different cationic lipids such as: dioleoyl trimethylammonium-propane (DOTAP), cetylpyridinium chloride, dimethyldioctadecylammonium bromide (DDAB), DC-cholesterol.

The cationic SLN are incubated with negatively charged nucleic acids; electrostatic interactions occur resulting in the formation of SLN-DNA or SLN-RNA complexes.

Positive charge of SLN-DNA/SLA-RNA surface is usually preferable for *in vitro* transfection, as it helps to promote the cellular uptake of the complex because of negative charge of cell membranes.

On the other side, use of some cationic lipids for *in vivo* delivery is limited by their toxicity [15].

The capacity of SLN-DNA vectors to in vivo transfect after intravenous administration in mice has been evaluated: as an example, cationic SLN were prepared by solvent emulsification evaporation method and by using DOTAP as cationic lipid, SLN-DNA vectors were obtained by mixing SLN with the plasmid pCMS-EGFP, which encodes the enhanced green fluorescent protein (EGFP). The intravenous administration in mice led to transfection in hepatic tissue and spleen, protein expression was detected from the third day after administration and it was maintained for at least 7 days; the results showed the capacity of SLN-DNA vectors to induce expression of foreign proteins in the spleen and in the liver assessing the potential of SLN for gene therapy [15].

SLN have been formulated also to incorporate nucleic acids inside their solid lipid matrix: this kind of nanoparticles are mainly prepared by (w/o/w) double warm microemulsion technique.

Efficacy of SLN incorporating vascular endothelial growth factor antisense oligonucleotide (VEGF-AS-ODN) to downregulate VEGF expression has been evaluated in rat glioma cells (in vitro) and in experimental murine model of glioma (in vivo) [2]: both experiments demonstrated that cellular VEGF expression was significantly reduced in tumor cells with SLN carrying VEGF-AS-ODN.

Oral Administration

SLN have been studied for oral administration of several drugs such as cyclosporin A, clozapine, tobramycin, idarubicin, and apomorphine.

SLN can improve oral bioavailability of drugs by different mechanisms [12]:

1. Protection of incorporated drugs from gastrointestinal fluids and enzymatic degradation
2. Transportation into lymphatic system avoiding first-pass metabolism
3. Absorption enhancing effect of lipids: lipids can promote the absorption of poorly water soluble active compounds
4. Bioadhesion of nanoparticles (due to their small size) to the intestinal wall increasing the residence time in intestinal tract

5. Effect of surfactants which may contribute to increase permeability of the intestinal membrane or to improve affinity between lipid particles and intestinal membrane

Gastrointestinal uptake of radiolabeled SLN administered to duodenum of rats by canula was studied [2]. SLN were observed in the lymph by using electron microscopy; the radioactivity data confirmed targeting of particles to lymph and blood.

In other in vivo study, in order to evaluate gastrointestinal absorption of drugs incorporated into SLN, tobramycin was selected as model drug because it is not absorbed by the gastrointestinal tract. SLN containing different percentages of tobramycin were administered in rats by canula to duodenum: the time-concentration curves showed different profiles (Fig. 8) [9] and pharmacokinetics parameters varied considerably among the three types of tobramycin-SLN. Possible reasons for the different behavior are number of SLN administered, particle size, total surface area, and drug concentration in each nanoparticle [9].

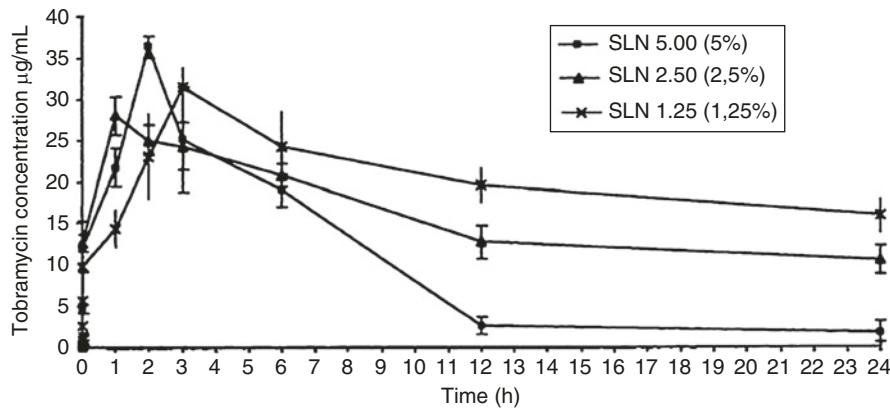
The results of this study confirmed transmucosal transport of drug when loaded in SLN to lymphatic system and the lymphatic uptake of tobramycin-SLN was detected also by TEM micrographs (Fig. 2) [9].

A critical parameter that has to be considered in oral administration is physicochemical stability of SLN into gastrointestinal fluids. SLN aggregation can occur in the stomach due to the acidity and high ionic strength of the gastric environment [1].

Drying of SLN dispersion into a dry powder (by freeze-drying and spray-drying process) can be necessary for the administration of solid dosage forms such as capsules and tablets.

Skin Application

SLN show many advantageous features for skin application. They can provide controlled release profiles, they are composed of physiological and biodegradable lipids with good tolerability, their small size allows a close contact with stratum corneum (horny layer) and causes formation of an adhesive film on the skin; film formation leads to an occlusive effect which increases skin



Solid Lipid Nanoparticles (SLN), Fig. 8 Tobramycin plasma concentrations versus time \pm SD after duodenal administration of Tobra-SLN, loaded at different percentage of tobramycin [9]

hydration, improving drug penetration into the skin. Furthermore, SLN can enhance chemical stability of compounds sensitive to light, oxidation, and hydrolysis [16].

The horny layer is an efficient barrier that protects humans from excessive water loss, toxic agents, and microorganisms: it is formed by corneocytes embedded into epidermal lipids forming highly structured layers. Below the horny layer there is the epidermis followed by the dermis.

In general lipid nanoparticles do not penetrate the horny layer, but a follicular uptake has been reported for particulate systems.

Incorporation of drugs into SLN can be exploited for both topical administration and for transdermal administration [2, 16].

SLN have been proposed for improvement of treatment of skin diseases (such as atopic eczema, psoriasis, acne, skin mycosis, and inflammations): drugs incorporated into SLN and investigated for dermal application have been glucocorticoids, retinoids, and antimycotics.

Experimental data showed that SLN can enhance drug penetration into the skin increasing treatment efficiency: improved bioavailability at the site of action reduces the required dose and reduces dose-dependent side effects of the drug. SLN can act as a drug reservoir and this can be an important tool when it is necessary to supply the drug over a prolonged period of time and when drug produces irritation in high concentration [16].

SLN have been investigated also as carriers in cosmetics, especially for UV blockers. SLN by themselves have a sun-protective effect, due to their solid particulate character, and are capable of reflecting UV radiation leading to photoprotection: molecular sunscreens showed to be much more effective after incorporation into SLN for synergistic effects [1, 16].

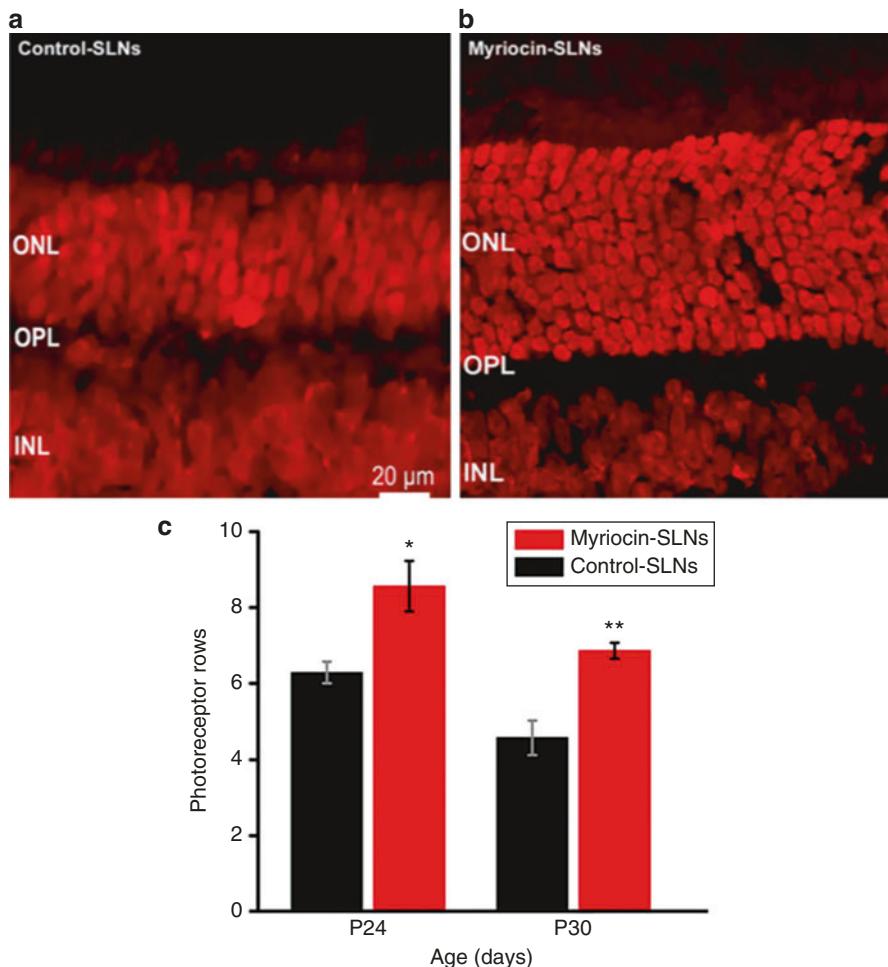
In order to allow skin application, SLN dispersion need to be incorporated into a cream or gel base which must not induce dissolution nor aggregation of SLN.

Ocular Topical Administration

Ocular bioavailability of drugs is limited by the complex structure of the eye: for ophthalmic application, drugs are usually formulated as eye-drops and administered topically but limited permeability of the cornea and other different mechanisms (such as lachrymal secretions, nasolacrimal drainage, drug adsorption into systemic circulation, drug spillage due to the limited capacity of human cul-de-sac, and drug metabolism caused by enzymes) strongly contribute to limit efficacy of the applied drug. Due also to this removal, several drug applications in a day are required to achieve therapeutic effect [17].

SLN can be useful system to enhance ocular bioavailability of both hydrophilic and hydrophobic drugs.

Possible advantages of SLN for improvement in ocular drug delivery include:

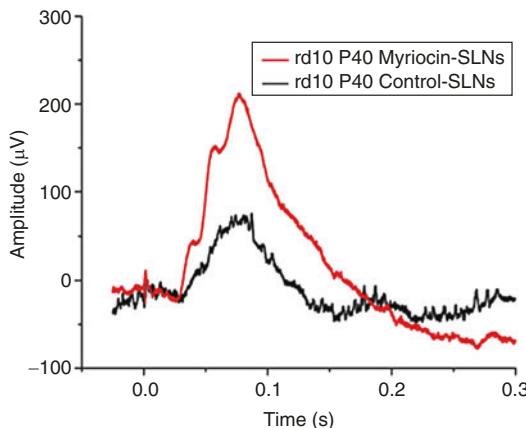


Solid Lipid Nanoparticles (SLN), Fig. 9 Effects of myriocin-SLN on retinal morphology: Vertical retinal sections from rd10 mice treated with control SLN (a) and myriocin-SLN (b) for 10 days (from P14 to P24). The outer nuclear layer (ONL) of the myriocin-treated retina

is thicker because it contains more photoreceptor rows than the control retina. INL inner nuclear layer, OPL outer plexiform layer. (c) Quantification of photoreceptor rows at P24 and P30 in rd10 mice treated with control SLN or myriocin-SLN [18]

1. Bioadhesive properties, which can prolong ocular surface residence time, with better drug penetration and reduced drainage, increasing drug ocular bioavailability and decreasing dosage and relevant side effects
2. Controlled release of drugs avoiding frequent administrations
3. Protection of incorporated drug from action of metabolic enzymes present on the ocular surface
4. Good tolerability as SLN can provide low irritation and good compatibility with ocular tissues
5. Eyedrops formulation, which is self-administered by patients, improving compliance

SLN have been investigated in animal model for ocular delivery of different drugs such as tobramycin, diclofenac, timolol, and cyclosporin A: specific SLN formulation allowed to assess



Solid Lipid Nanoparticles (SLN), Fig. 10 Cone-driven ERG (Electroretinography) responses from rd10 mice (red trace) treated with myriocin-SLN or control SLN (black trace). Animals were age P40 and treatment had initiated at P14 [18]

efficacy of myriocin (MYR) in retinitis pigmentosa animal model [18].

MYR is a hydrophobic drug, it can inhibit intracellular synthesis of ceramide, a physiological lipid involved in apoptotic cascade. MYR has been proposed as drug for treatment of retinitis pigmentosa (RP), a disease where degeneration of photoreceptor causes blindness: there is not efficacious treatment approved for RP at this time. Efficacy of MYR loaded SLN has been showed in RP animal model (RD10 mouse model): three times a day administration by eyedrops of MYR-SLN, lasting a period of 35 days, did not show toxicity in mice and physiological registration of activity of photoreceptors by electroretinography (ERG) showed reduced degeneration occurred in animals treated with MYR-SLN (Figs. 9 and 10) [18].

Pulmonary Administration

The large inner surface of the lung and the thin alveolar epithelium allow rapid drug absorption avoiding first-pass metabolism; pulmonary administration using solid lipid nanoparticles represents an alternative for both local and systemic drug delivery.

SLN show different advantages for pulmonary drug delivery, such as the possibility to have controlled release profile, high tolerability, and possibility to reach lymphatic system through the action of alveolar macrophages. For pulmonary administration, SLN dispersions can be nebulized; inhalation device (design of nebulizer, flow rate) and physicochemical properties of formulations can affect aerosol nebulization efficiency and site of aerosol deposition.

Toxicological studies have been performed by using in vitro and in vivo tests [19]: A549 cells and murine precision-cut lung slices (PCLS) were exposed to increasing concentrations of SLN in order to estimate the toxic dose of SLN. The in vitro experiments showed toxic effects begin at concentrations of about 500 µg/mL, while for in vivo experiments toxicological potential of SLN was determined in a 16-days repeated dose inhalation study using mice which were daily exposed to different concentration of SLN (1–200 µg deposit dose): results showed that repeated inhalation exposure to SLN is safe in murine inhalation model.

Biodistribution of inhaled radiolabeled SLN in rats was studied [20]. Results show an important and significant SLN uptake into the lymph, few minutes after inhalation SLN translocate and accumulate into regional lymph nodes suggesting translocation mechanism of SLN may involve phagocytosis by bronchoalveolar macrophages followed by migration to lymphatic system.

S

Cross-References

- [Liposomes](#)
- [Nanoencapsulation](#)
- [Nanomedicine](#)
- [Nanoparticle Cytotoxicity](#)
- [Nanoparticles](#)
- [RNAi in Biomedicine and Drug Delivery](#)

References

1. Mehner, W., Mäder, K.: Solid lipid nanoparticles: production, characterization and applications. *Adv. Drug Deliv. Rev.* **47**, 165–196 (2001)
2. Gasco, M.R., Mauro, A., Zara, G.P.: In vivo evaluations of solid lipid nanoparticles and microemulsions. In: *Drug Delivery Nanoparticles Formulation and Characterization*, pp. 219–238. Informa Healthcare, New York (2009)
3. Manjunath, K., Venkateswarlu, V.: Pharmacokinetics, tissue distribution and bioavailability of clozapine solid lipid nanoparticles after intravenous and intraduodenal administration. *J. Control. Release* **107**, 215–228 (2005)
4. Siekmann, B., Westesen, K.: Investigations on solid lipid nanoparticles prepared by precipitation in o/w emulsions. *Eur. J. Pharm. Biopharm.* **43**, 104–109 (1996)
5. Quintanar-Guerrero, D., Tamayo-Esquivel, D., Ganem-Quintanar, A., Allemann, E., Doelker, E.: Adaptation and optimization of the emulsification-diffusion technique to prepare lipidic nanospheres. *Eur. J. Pharm. Sci.* **26**, 211–218 (2005)
6. Schubert, M.A., Müller-Goymann, C.C.: Solvent injection as a new approach for manufacturing lipid nanoparticles – evaluation of the method and process parameters. *Eur. J. Pharm. Biopharm.* **55**, 125–131 (2003)
7. Battaglia, L., Gallarate, M., Cavalli, R., Trotta, M.: Solid lipid nanoparticles produced through a coacervation method. *J. Microencapsul.* **27**, 78–85 (2010)
8. Charcosset, C., El-Harati, A., Fessi, H.: Preparation of solid lipid nanoparticles using a membrane contactor. *J. Control. Release* **108**, 112–120 (2005)
9. Cavalli, R., Bargoni, A., Podio, V., Muntoni, E., Zara, G.P., Gasco, M.R.: Duodenal administration of solid lipid nanoparticles with different percentages of tobramycin. *J. Pharm. Sci.* **92**, 1085–1094 (2003)
10. Yuan, H., Huang, L.-F., Du, Y.Z., Ying, X.Y., You, J., Hu, F.Q., Zeng, S.: Solid lipid nanoparticles prepared by solvent diffusion method in a nanoreactor system. *Colloids Surf. B Biointerfaces* **61**, 132–137 (2008)
11. Bunjes, H.: Lipid nanoparticles for the delivery of poorly water-soluble drugs. *J. Pharm. Pharmacol.* **62**, 1637–1645 (2010)
12. Muchow, M., Maincent, P., Muller, R.H.: Lipid nanoparticles with a solid matrix (SLN, NLC, LDC) for oral drug delivery. *Drug Dev. Ind. Pharm.* **34**, 1394–1405 (2008)
13. Joshi, M.D., Muller, R.H.: Lipid nanoparticles for parenteral delivery of actives. *Eur. J. Pharm. Biopharm.* **71**, 161–172 (2009)
14. Bondi, M.L., Craparo, E.F.: Solid lipid nanoparticles for applications in gene therapy: a review of the state of the art. *Expert Opin. Drug Deliv.* **7**, 7–18 (2010)
15. Del Pozo-Rodriguez, A., Delgado, D., Solinis, M.A., Pedraz, J.L., Echevarria, E., Rodriguez, J.M., Gascon, A.R.: Solid lipid nanoparticles as potential tools for gene therapy: in vivo protein expression after intravenous administration. *Int. J. Pharm.* **385**, 157–162 (2010)
16. Pardeike, J., Hommoss, A., Müller, R.H.: Lipid nanoparticles (SLN, NLC) in cosmetic and pharmaceutical dermal products. *Int. J. Pharm.* **366**, 170–184 (2009)
17. Seyfoddin, A., Shaw, J., Al-Kassas, R.: Solid lipid nanoparticles for ocular drug delivery. *Drug Deliv.* **17**, 467–489 (2010)
18. Strettoi, E., Gargini, C., Novelli, E., Sala, G., Ilaria, P., Gasco, P., Ghidoni, R.: Inhibition of ceramide biosynthesis preserves photoreceptor structure and function in a mouse model of retinitis pigmentosa. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 18706–18711 (2010)
19. Nassimi, M., Schleh, C., Lauenstein, H.D., Hussein, R., Hoymann, H.G., Koch, W., Pohlmann, G., Krug, N., Sewald, K., Rittinghausen, S., Braun, A., Müller-Goymann, C.: A toxicological evaluation of inhaled solid lipid nanoparticles used as potential drug delivery system for the lung. *Eur. J. Pharm. Biopharm.* **75**, 107–116 (2010)
20. Videira, M.A., Botelho, M.F., Santos, A.C., Gouveia, L.F., De Lima, J.J., Almeida, A.J.: Lymphatic uptake of pulmonary delivered radiolabelled solid lipid nanoparticles. *J. Drug Target.* **10**, 607–613 (2002)

Solid Lipid Nanospheres

► [Solid Lipid Nanoparticles \(SLN\)](#)

Solid Lipid-Based Nanoparticles

► [Solid Lipid Nanoparticles \(SLN\)](#)

Solid-Liquid Interfaces

► [Nanoscale Properties of Solid-Liquid Interfaces](#)

Solid-State Heat Convertors

Joseph P. Heremans

Department of Mechanical and Aerospace Engineering, Department of Physics and Department of Materials Science and Engineering, The Ohio State University, Columbus, OH, USA

Definition

Solid-state Heat Converters are solid-state devices that convert heat into electrical work without any moving parts, for example, by using thermoelectric effects or heat-induced changes in magnetic effects.

Motivation

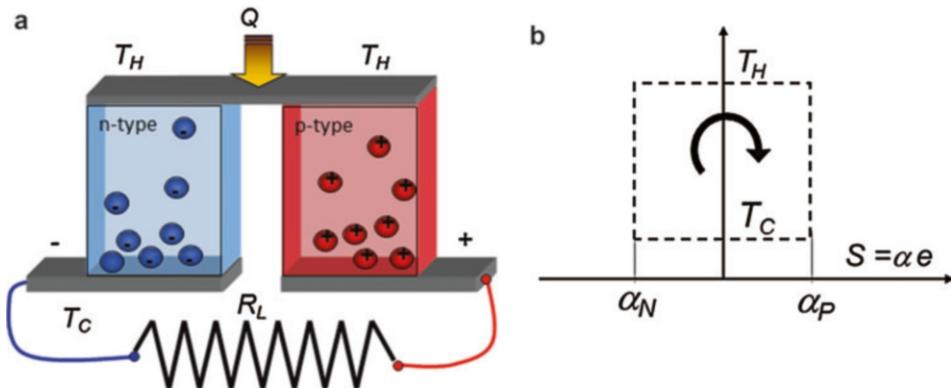
Thermoelectric energy conversion emerged in the middle of the twentieth century. Early pioneers were Maria Telkes [1, 2], Abram Ioffe [3], and Julian Goldsmid [4], who developed thermoelectric semiconductors and showed that they could be used for electrical power generation from heat and for solid-state cooling. Our understanding of thermoelectricity itself dates back almost two centuries, with the discovery of the thermoelectric power of solids (the Seebeck coefficient) by Thomas Seebeck [5] in 1821–1822. Lord Kelvin [6] discovers the Thomson heat and understood thermoelectric effects to be part of classical thermodynamics. He derived his reciprocity relations by treating electrical currents like fluids in conventional mechanical heat engines. Lord Rayleigh [7] realized that thermoelectric effects could theoretically be used for solid-state heat-to-electricity conversion. He pointed out not only that the efficiency of thermopiles is proportional to the square of the thermopower, but also that “...the steam-engine and dynamo are not likely to be superseded by German-silver (a Cu-Ni alloy related to constantan) and iron thermopiles.” Altenkirch [8]

established the concept of the thermoelectric figure of merit ZT , which will be described further in these pages. In practice, before 1945 the only use of thermoelectric effects was in metallic thermocouples used for thermometry.

When a solid is subjected to a temperature gradient (∇T), a heat flux (j_Q) flows through it. To the first order, the heat flux is linearly related to ∇T via Fourier’s law, $j_Q = \kappa \nabla T$, where κ is the solid’s thermal conductivity. Both ∇T and j_Q are vectors; in the absence of external magnetic fields, they are collinear. When the solid is electrically conducting (metals or semiconductors), ∇T also generates an electric field E , again aligned with ∇T in the absence of an external magnetic field. In linear transport theory, both are again proportional and the thermoelectric power α of the solid, also known as thermopower or Seebeck coefficient, is the proportionality constant, defined as $\alpha \equiv E / \nabla T$. It has been shown by Callen [9] that, for a single electron and in strictly reversible thermodynamics, the thermopower equals the entropy S of that electron divided by its charge e , $\alpha = S/e$. Consequently:

1. The thermopower is a state equation: it does not matter how the electron goes from point A to B in a sample, just what the potential and temperature are at points A and B.
2. The thermopower is an absolute property of a material, not a difference. It does not require a thermocouple made from two materials to measure a thermopower. The absolute value of thermopower is measured by calorimetry by integrating the Thomson heat [10] over temperature.
3. The Nernst principle (third law of thermodynamics) holds to the thermopower: in the limit for zero absolute temperature, the entropy and therefore the thermopower are zero.
4. One can build heat engines with thermopower, either to generate power from heat or to operate as cooling/refrigeration cycles.

Figure 1a shows a thermoelectric power generator, which consists of two thermoelectric



Solid-State Heat Convertors, Fig. 1 (a) Schematic diagram of a thermoelectric power generator consisting of an n-type semiconductor and a p-type semiconductor connected as shown. (b) The thermodynamic cycle an

semiconductors: one is doped n-type in which the majority carriers are electrons, and the thermopower is negative ($\alpha_N < 0$); and other one is doped p-type in which the majority carriers are holes, and the thermopower $\alpha_P > 0$. These semiconductor elements are connected in a thermopile, with a hot contact at temperature T_H and a cold contact at temperature T_C ; the difference in temperature between the contacts is maintained by adding heat Q to the hot contact. The electrical contacts are connects in series, as shown in Fig. 1a: an electrical load R_L is attached on the cold side between the n- and p-type semiconductors. The temperature difference $T_H - T_C$ creates a voltage between the two bottom plates, which, in open circuit conditions (i.e., if $R_L \rightarrow \infty$), would be equal to $(\alpha_P - \alpha_N)(T_H - T_C)$ with the polarity shown in Fig. 1a. Assume further that α_p and α_N are not temperature-dependent between T_H and T_C . Looking at Fig. 1a from the point of view of one electron, starting from the bottom left (cold side, n-type), as that electron rises in the structure, its temperature increases at constant value of α_N , i.e., isentropically at entropy $S_N = \alpha_N e$, under Callen's condition that the process were reversible. That progress can be followed in the (T,S) diagram (Fig. 1b). When the electron has reached the top hot plate, it moves over from left to right isothermally at T_H , as can be followed in Fig. 1b. It then cools back down to T_C in the p-type

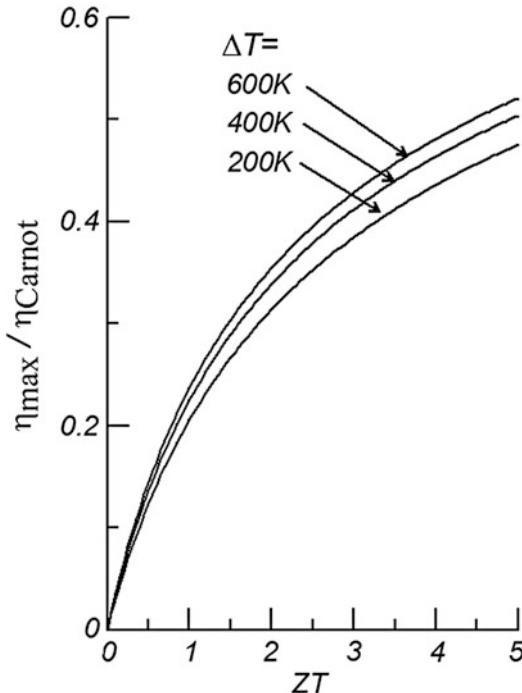
electron goes through as it travels through the power generator, represented in a (temperature, entropy) diagram which is the equivalent of a (temperature, thermopower) diagram

semiconductor, another isentropic transformation, at a constant entropy $S_P = \alpha_P e$, again assuming the process is considered reversible. Finally, it flows through the load R_L isothermally again at the cold temperature T_C . Therefore, in reversible thermodynamics, a thermoelectric power generator as shown in Fig. 1a would have a thermodynamic cycle as shown in Fig. 1b, which is a Carnot cycle, and the efficiency would be the Carnot efficiency

$$\eta = \eta_C = \frac{T_H - T_C}{T_H}. \quad (1)$$

Peltier coolers, under the same assumption, would also have the Carnot Coefficient of Performance. Clearly, none of these assertions are true, because the losses in thermoelectric generator and Peltier cooler are dominated by two main irreversibilities:

1. Heat is lost by thermal conduction in the semiconductors. This loss is quantified by the thermal conductivity of the materials (κ_N and κ_P).
2. Joule heating is also an irreversible process. It is due to the internal resistance of the thermopile though which the electrical current must pass in order to do electrical work in the load. Joule heating is minimized by maximizing the electrical conductivity σ of each of the semiconductors.



Solid-State Heat Convertors, Fig. 2 Thermal efficiency of the thermoelectric generator of Fig. 1, as a function of the average ZT of the n- and p-type semiconductors, for the values of temperature difference (ΔT) indicated. The efficiency is represented as a function of Carnot efficiency and includes the irreversible thermodynamic losses of the device

It is Altenkirch [7] who first quantified the role of these irreversibilities and related the efficiency of the thermoelectric generator in Fig. 1 to the thermoelectric figure of merit ZT of each of the two semiconductors. This is defined by:

$$ZT \equiv \frac{\alpha^2 \sigma}{\kappa} T. \quad (2)$$

The ZT must be characterized separately for the n- and p-type materials, but for simplicity we will assume that they are equal and omit the indices (or, to a first approximation, assume that for the device ZT is the average between ZT_N and ZT_P). If so, the efficiency is shown as a function of ZT in Fig. 2. We also define the thermoelectric power factor PF :

$$PF \equiv \alpha^2 \sigma. \quad (3)$$

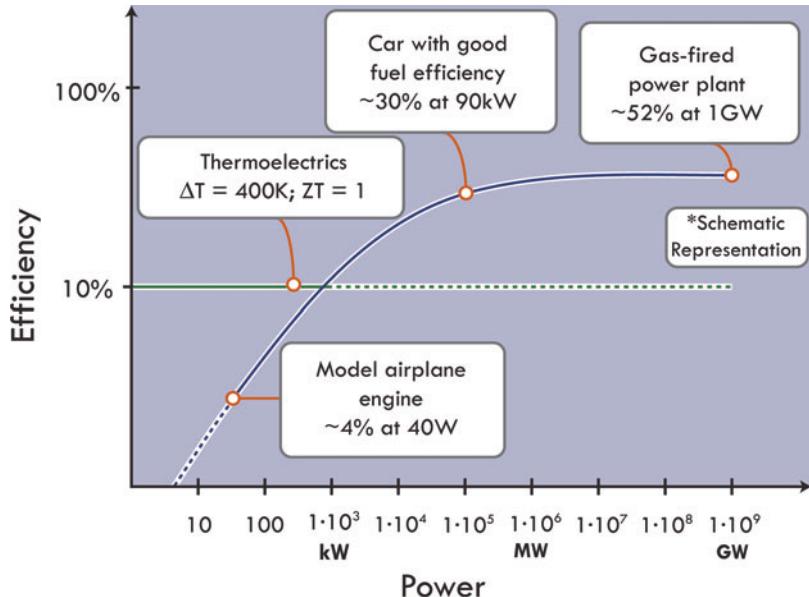
The PF , which appears in the numerator of ZT , is a measure of how much electrical power per unit volume of material a generator can produce under a constant temperature difference. Indeed, under a given $(T_H - T_C)$, α determines the voltage, $\alpha \sigma$ the current density, and the product of the two the power density.

The TE generators such as shown in Fig. 1 have many practical advantages over conventional internal combustion engines or steam-power cycles. They have no moving parts, do not wear out, require no maintenance, are extremely reliable, and have almost infinite lifetimes. These generators have been used to produce power in interplanetary space probes (where the source of heat are radioactive isotopes) where they have operated continuously for more than half a century. They also have extremely high power densities (the number of Watts per unit volume of material), which implies that they scale well to low power densities (1 W or less). The main drawback of TE generators is that at large power levels they are far less efficient than conventional heat engines [11], because the ZT of today's materials is too low. This relative disadvantage turns into an advantage in low-power applications, because conventional fluid-based machines lose their efficiency with power level. Figure 3 shows that there is a power range (< 1 kW, perhaps < 100 W) where TE generators compete well [12] with mechanical fluid-based engines. Nevertheless, in the last 20 years, a significant effort is devoted to finding ways to improve the ZT other TE semiconductors, mainly based on the introduction of nanotechnologies [12].

The main difficulty in thermoelectric research is that α , σ , and κ are interrelated and almost mutually counter-indicated in the way they enter (Eq. 2). Yet, they have to be optimized on the same material. To illustrate this, one can re-write the ZT recalling that the electrical conductivity is a function of the density of charge carriers n and their mobility μ , and by recalling that the thermal conductivity is due to contributions of κ_E from the electrons and κ_L from the lattice vibrations or phonons:

Solid-State Heat Convertors,

Fig. 3 Schematic comparison of the efficiency of conventional heat engines and thermoelectric converters as a function of power level. Thermoelectric solid-state heat engines are competitive in only low-power applications



$$ZT = (\alpha^2 n) \left(\frac{\mu}{\kappa_E + \kappa_L} \right) eT. \quad (4)$$

$$\kappa_E = LT\sigma. \quad (5)$$

The factor $(\alpha^2 n)$ in Eq. 4 contains two mutually counter-indicated properties α and n . Indeed, the thermoelectric power decreases as the charge carrier concentration increases in most solids, a relation attributed to Mr. Pisarenko [13] and shown Fig. 4. As a result, the power factor has a maximum at the optimal charge carrier concentration n_{OPT} , typically between $1 \times 10^{18} \text{ cm}^{-3}$ for solids with an electron effective mass m^* below $0.1 m_e$ (the free electron mass) to $5 \times 10^{20} \text{ cm}^{-3}$ for solids with $m^* > m_e$. In the more common thermoelectric semiconductors with $0.25 m_e \leq m^* \leq 0.7 m_e$, we have $2 \times 10^{19} \text{ cm}^{-3} \leq n_{OPT} \leq 7 \times 10^{19} \text{ cm}^{-3}$. Further, the higher the value of the density of states effective mass m^* , the higher the value of the optimum power factor. This illustrates the best way maximize $(\alpha^2 n)$: select solids with high effective masses.

The factor $\mu/(\kappa_E + \kappa_L)$ in Eq. 4 contains another two other mutually counter-indicated properties μ and κ_L . Indeed, adding defects to the system to decrease the thermal conductivities of either electrons or phonons usually also decreases the electron mobility. To illustrate how to maximize $\mu/(\kappa_E + \kappa_L)$, we re-writing Eq. 4 making use of the Wiedemann-Franz law that relates κ_E and σ .

Where L is the Lorentz ratio. For free electrons, $L = L_0$, a smattering of universal constants giving $L_0 \equiv 2.4 \times 10^{-8} \text{ V}^2 \text{ K}^{-2}$. In practice, L varies little from L_0 , and is typically $0.6 L_0 < L < 1.4 L_0$. Substituting this in Eq. 2 gives:

$$ZT = \frac{\alpha^2}{L} \frac{T}{1 + \frac{\kappa_L}{\kappa_E}}. \quad (6)$$

This illustrates that thermoelectrics research must aim at reducing the lattice thermal conductivity, not necessarily in absolute numbers, but with regards to the electronic thermal conductivity. It also illustrates the importance of maximizing the thermopower α as long as the electrical conductivity is high enough. In the extreme case of very highly conducting metals, where the fraction of the heat carried by the phonons is zero, $ZT = \frac{\alpha^2}{L}$, and the primary goal of research has now shifted from optimizing ZT to optimizing a single transport property, the thermoelectric power α .

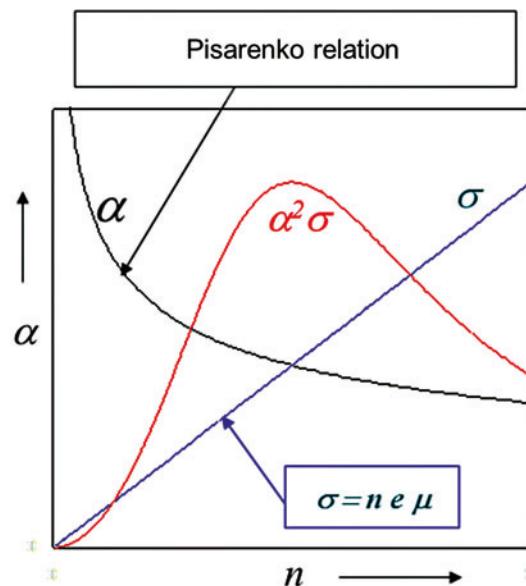
In the light of the above discussion, the following strategies are used to optimize ZT .

1. Phonon engineering to reduce lattice thermal conductivity
 - 1a. Engineering nanostructures that scatter phonons more than electrons
 - 1b. Engineering localized phonon modes (rattlers) to scatter acoustic phonons
 - 1c. Engineering and maximizing the anharmonicity of the chemical bonds to promote phonon-phonon interactions
2. Band structure engineering to enhance thermopower $\alpha(n)$
 - 2a. Engineering nanostructures to size-quantize electrons, using two-dimensional structures, quantum wells, quantum wires, and quantum dots.
 - 2b. Engineering resonant impurities that produce similar effects in the band structure, but now in bulk solids.
 - 2c. Engineering spin polarization in Kondo physics or correlated systems
3. Add an additional optimization parameter to mitigate the problem of having to optimize counter-indicated properties of a single solid: magnetism and spin physics

From Eq. 6 and the discussion above, it is also clear that reducing the relative role of the lattice thermal conductivity can be achieved by either reducing κ_L itself or increasing σ and κ_E as long as α is not affected: it is thus more effective to increase the power factor than to decrease the thermal conductivity. These strategies are discussed below.

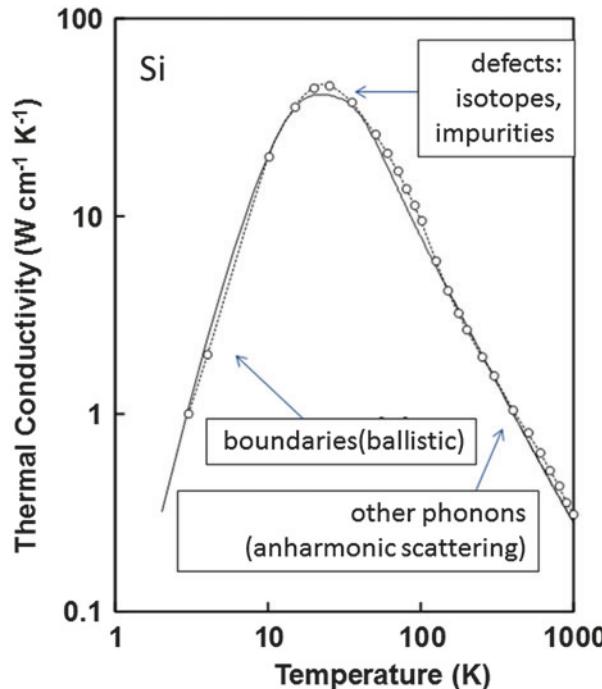
Phonons

Lattice thermal conductivity κ_L is the only heat-carrying mechanism in nonmagnetic dielectric solids, and, in crystals, it is mainly due to the contribution of the acoustic phonons to the thermal conductivity [14]. These phonons are scattered by either other phonons (phonon-phonon interactions) or various defects. The temperature dependence of the thermal conductivity of high-purity single-crystal silicon [15], shown in Fig. 5, illustrates the various regimes. The defects are generally classified by their size in relation to



Solid-State Heat Convertors, Fig. 4 Schematic relation between thermopower (α), electrical conductivity (σ), and thermoelectric power factor ($\alpha^2 \sigma$) and the concentration n of charge carriers (electrons or holes)

the wavelength of the phonons they interact with. The phonons that carry heat are dominantly acoustic phonons, which come in three modes, one longitudinal (LA) and two transverse (TA) acoustic modes. Each mode has a dispersion at temperatures much below the Debye temperature Θ_D of the solid given by $E = \hbar\omega = \hbar k v_G$, where v_G is the phonon mode's group velocity (the sound velocity); k is the phonon wave vector, the inverse of its wavelength $\Lambda = 1/k$. Because the average phonon energy E is on order of $k_B T$, the above equation immediately shows the inverse relation $k_B T = \hbar v_G / \Lambda$ between temperature and wavelength: the abscissa axis of Fig. 4 therefore can be seen as proportional to the inverse wavelength of the phonons of average energy at that given temperature, called “thermal phonons.” There is an intuitive direct relation between wavelength and mean free path of sound waves: longer wavelength phonons are scattered by larger-scale obstacles. Therefore, Fig. 4 is explained in terms of the length scales of the obstacles that scatter acoustic phonons.



Solid-State Heat Convertors, Fig. 5 Thermal conductivity of single-crystal silicon as a function of temperature. The thermal conductivity is entirely due to phonons. Below 10 K it is limited by interactions between phonons and the sample boundaries, i.e., transport is ballistic. Above 80 K it is limited by interactions between phonons, which are

governed by the anharmonic nature of the bonds of the crystal. In the intermediate temperature range, the thermal conductivity has a maximum, the value of which is limited by interactions with phonons with various defects, such as various isotopes of Si or impurities or dislocations

Ballistic Phonons

The amount of heat carried by each mode is given by the kinetic formula, where phonons are considered quasi-particles that behave like an ideal gas: $\kappa_L = \frac{1}{3} C v_G \ell$. Here the mean free path $\ell = v_G \tau$ is expressed as a function of the group velocity and the scattering time τ , which is the inverse of the scattering rate τ^{-1} . The scattering rate depends on the nature of the obstacle that scatters the phonon. At low temperature (left, in Fig. 5), where phonon wavelengths are very long, the phonon mean free path can reach millimeters at cryogenic temperatures and ℓ is limited by the crystal boundaries. Phonon transport becomes ballistic. It is characterized by a scattering time τ_B , calculated for a crystal whose smallest size is defined as t to be $\ell = t$, $\tau_B = t/v_G$. To reduce κ_L in this temperature range, good thermoelectrics are prepared as fine-grained crystals, typically $t < 1 \mu\text{m}$ and on the order of 100 nm, by ball-

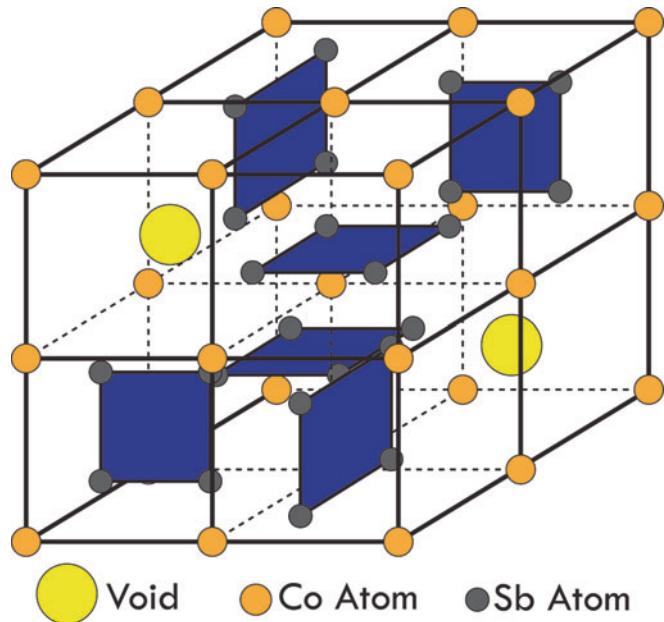
milling and sintering [16]. This has led to important reductions in κ_L and increases in the ZT of the most widely used thermoelectrics for Peltier cooling at room temperature, p-type alloys of approximate formulas $(\text{Bi}_{0.25}\text{Sb}_{0.75})_2\text{Te}_3$, and n-type alloys of a composition near $\text{Bi}_2(\text{Se}_{0.1}\text{Te}_{0.9})_3$ [16].

Phonon-Phonon Interactions

At a temperature on the order of $T \leq \Theta_D/3$, the phonon wavelengths become of the order of one or a few interatomic distances or crystal unit cells, and phonon-phonon scattering (with a scattering probability τ_ϕ^{-1}) dominates. The scattering frequency, τ_ϕ^{-1} , is the product of three factors: (1) the number of phonons available for scattering (i.e., the phonon concentration), (2) the number of empty states available for phonons to scatter into, and (3) the probability of the interaction of each phonon with

Solid-State Heat

Convertors, Fig. 6 The structure of the unfilled skutterudite CoSb_3



another phonon. All three of these properties can be rigorously evaluated, but they involve nested integrals over the entire Brillouin zone of all phonon modes [17]. Solids can be selected that maximize any of these three properties and have minimum lattice thermal conductivity; some solids can even be engineered to induce these phonon properties.

Selecting solids that have a low density of states in acoustic modes is typically done by using large unit cells. Indeed, this implies that the reciprocal cell and the Brillouin zone are small and have a low density of states. Many good thermoelectrics have long and complex chemical formulae, which achieves this goal. A corollary is that what would otherwise be acoustic modes in solids with simple unit cells become optical modes in more complex solids built from a multiplicity of such simple unit cells. Optical modes not only have low group velocities, and therefore do not carry heat efficiently, but they are also very helpful with the next mechanism.

Selecting solids with a large number of available phonon modes for acoustic phonons to scatter into generally is done by selecting solids with high densities of optical modes, as described

above, with energies coinciding with the energies of the acoustic phonons. This approach has been pushed into phonon-mode engineering, typically applied to a family of high-temperature thermoelectrics of general formula CoSb_3 that belongs to the crystallographic family called skutterudites after a natural mineral $(\text{Co}, \text{Ni}, \text{Fe})\text{As}_3$ ore. While n-type CoSb_3 has a good power factor, its lattice thermal conductivity is too high to have a good ZT . The unit cell of the crystal consists of eight elementary cubes with Co atoms on their corners, as shown in Fig. 6. Two of these cubes are empty, and the other six are filled with planar rings of Sb, arranged so as to form octahedra with a Co atom in the middle. The empty cubes act as voids that can accept an electronegative atom in the middle, creating “filled” skutterudites. When this atom is small, it can “rattle” in the void, creating an optical phonon mode that can accept energy from the heat-carrying acoustic modes, dramatically lowering the lattice thermal conductivity [18–20]. This effect has resulted in alloy compositions for high-temperature thermoelectrics that have very high ZT values reaching $ZT \approx 1.7$ at 800 K [21].

Selecting solids in which the phonon-phonon interactions are maximized has also been achieved

recently. Returning to the temperature dependence of κ_L at $T > 100$ K in Fig. 4, one observes that it scales roughly as a T^{-1} law, which is characteristic for a conductivity limited by phonon-phonon scattering. A very simplified expression is given by Berman [14], valid for temperatures of about $T > \Theta_D/2$:

$$\kappa_L = A \frac{\bar{M} \Theta_D^3 V_{\text{atom}}^{1/3}}{\gamma^2 n^{\gamma/3} T} \quad (7)$$

Here, \bar{M} is the average mass of atoms, V_{atom} the volume per atom, n the number of atoms in the unit cell, A a collection of universal constants (3×10^{-5} if SI units are used consistently), and γ is the average Grüneisen constant. The Grüneisen constant is a mode- and frequency-dependent property that characterizes the change of the frequency ω of a given phonon with the volume V , $\gamma(\omega) \equiv d\ln(\omega)/d\ln(V)$ and is mapped out across the Brillouin zone. The Grüneisen constant arises from the anharmonicity of the chemical bond: if the interatomic bonds behaved like perfect harmonic oscillators, or like balls and springs with position-independent spring constants, $\gamma = 0$. The fact that the spring constants do depend the atomic displacements (via V), i.e., the anharmonicity, is reflected on the acoustic properties of the medium and phonon eigenfrequencies ω . The effect of γ on the probability of phonon-phonon interactions arises as follows. In an anharmonic solid, when a first phonon passes, the atomic displacements involved change the local values of ω , thereby modifying the acoustic property of the medium for a second phonon, should one pass by. The second phonon is thereby more likely to be reflected by the first phonon in a solid with a higher γ ; the probability of interaction τ_ϕ^{-1} therefore scales with γ^2 , which is reflected in Eq. 7.

How does one select solids with large anharmonicities? Solids where the atoms have high coordination numbers (octahedral bonds or more) are more anharmonic than those with lower numbers (tetrahedral bonds), favoring rocksalt (PbTe) or tetradymite (Bi_2Te_3) structures over diamond (Si) or wurtzite (ZnS) structures. Recent

progress has started from identifying the role of the lone pair electrons [22] in compounds of the class of the I-VI₂ compounds. Here, I represents a group I element: an alkali or a noble metal atom; V represents a pnictogen atom: As, Sb, or Bo; and VI represents a chalcogen atom: S, Se, or Te. These I-VI₂ compounds crystallize in a cubic inverted CaCl₂ structure. The paradigm for such materials, AgSbTe₂, has a ZT exceeding 1 [23]. While one reason is its very favorable band structure and power factor [24], the main effect is due to a lattice thermal conductivity that is pinned at the amorphous limit by phonon-phonon interactions [25] due to the extremely high γ . It was later shown that this is a general property of all rocksalt I-VI₂ compounds [22], due to the extreme sensitivity of the valence electrons whose orbital nature corresponds to the s-electrons of the group V element (the “lone pair” electrons in chemical terms) to heat-carrying acoustic phonons, resulting in an extremely high value of γ for those specific phonons.

There are several other structures where similar anharmonicities are at work, although not necessarily on the same phonons or due to the same parts of the valence band structure. One example involves the copper atoms in Cu-Sb-Se tetrahedrites [26], a naturally occurring mineral that, with some chemical modifications, can be made into an excellent thermoelectric. The role of lone pair electrons on the pnictogen atom in the thermal conductivity of tetradymites (Bi_2Te_3 , Bi_2Se_3 , Sb_2Te_3) has not yet been studied. In contrast, the role of Pb and Sn on that in PbTe is extensively investigated. It is proven that PbTe hosts some extraordinarily strong interactions between acoustic and optical phonons [27], an effect that contributes much to the very high ZT of PbTe. The interactions in the lead and tin salts involve optical phonons. Thus, they are subtly different from those that affect the acoustic phonons of the I-VI₂ compounds. However, such considerations in no way detract from the fact that anharmonicity is the key to the high ZT of the lead salts and of the record-holding $ZT = 2.6$ obtained in SnSe [28]. This result is made more remarkable by the fact that it was obtained on single crystals of a simple binary compound, without any alloying or

nanostructuring [29], and must therefore be intrinsic, such as only phonon-phonon interactions are known to be.

Nanostructures

At an intermediate temperature scale on the order of $\Theta_D/10$ to $\Theta_D/2$, the lattice thermal conductivity of very pure large single crystals shows a maximum (see Fig. 5) where boundary scattering and phonon-phonon interactions are both low. Since ℓ of the thermal phonons is very long at the temperature of this maximum, any extrinsic defects remaining in the sample will scatter the phonons. The phonon wavelength here is on the order of $\Lambda \sim 100$ nm to 10 μm , and defects that can limit ℓ at these temperatures are those at those at a length scale about ten times smaller, namely a few to a few hundred lattice unit cells. Some defects are unavoidable, such as naturally occurring isotopes. Dislocations and foreign atoms are other impurities that scatter phonons most prominently here; they have characteristic frequency and temperature dependences [14], but often these impurities tend to scatter electrons as much as phonons and do not increase ZT efficiently. It is here that nanotechnologies have had their strongest influence on reducing κ_L in thermoelectric semiconductors.

There are two ways to create nanostructures in thermoelectric semiconductors, metallurgical heat treatments and powder metallurgy. Metallurgical heat treatments can be used to create nanoprecipitates of a second phase in compound semiconductors of the parent phase, mimicking the formation of perlitic steels or Guinier-Preston zones in duralumin. Early on, excess Pb nanoparticles were precipitated in PbTe [30]. In order to improve ZT values, the precipitates must let electrons through but scatter phonons, thereby improving the $\mu/(\kappa_E + \kappa_L)$ ratio in Eq. 4. Kanatzidis and his group discovered that the best way to achieve this is to find secondary phases in the thermoelectric semiconductors that are also semiconductors but have electronic bands that align with those of the host material. These secondary phases should have very similar crystal structures and grow in crystallographic registry (“endotaxially”). Examples are SrTe

nanoparticles in PbTe and also PbS in PbTe. PbS within the PbTe matrix further beneficially modifies the density of electron states (see below) to allow for enhancement of the power factor to give a ZT of 1.8 at 800 K [31]. Other groups have used this technique successfully on various other materials.

Powder metallurgy can be used to create nanoinclusions, or at least crystals of nanometer-sized grains, that will scatter intermediate-energy phonons when made at the 10–100 nm scale, as mentioned in the section on **ballistic phonons**. This method was used successfully to prepare tetradymite materials that hold the record for ZT in their temperature ranges [16].

All-Scale Hierarchical Phonon Scattering

The record value of ZT , prior to the value reached on SnSe, is obtained on PbTe that includes nano-scale SrTe. It is prepared by ball-milling and spark plasma sintering and reaches $ZT = 2.2$ at 900 K [32]. This material, in essence, combines all of the techniques described above to reduce κ_L without affecting the mobility too much. The SrTe nanostructures scatter phonons with short and medium mean free paths (3–100 nm), but leave phonons with longer mean free paths unaffected. The phonons with longer mean free paths (0.1–1 mm) are affected by the grain-boundary phonon scattering. At the highest temperature, the strong anharmonicity promotes the phonon-phonon interactions that scatter the shortest-wavelength phonons (<3 nm), an intrinsic mechanism that is present in all PbTe samples. The authors label this the all-scale hierarchical architecture for designing thermoelectric materials.

S

Electrons

Hicks and Dresselhaus pointed out in two seminal papers [33, 34] in 1993 how quantum size effects could open a new way to design thermoelectric materials and had the potential to increase the ZT of narrow-gap semiconductors and semimetals to values of up to 5. The basic idea, which was subsequently fleshed out by Mahan and Sofo [35], stems from the Mott relation [36] between

the thermopower and the energy-dependent “Mott” conductivity $\sigma(E)$, a quantity which, when integrated over energy, gives the electrical conductivity $\sigma = -\int \sigma(E) (\partial f / \partial E) dE$. The Mott relation is:

$$\alpha = \frac{k_B}{e\sigma} \int \sigma(E) \frac{E - E_F}{k_B T} \left(\frac{\partial f}{\partial E} \right) dE. \quad (8)$$

Here, f is the Fermi-Dirac distribution function and E_F the Fermi energy. For non-degenerately doped semiconductors, the Mott relation yields the Pisarenko relation mentioned in section “Motivation.” For metals and degenerately doped semiconductors, the Bethe-Sommerfeld expansion (a second-order Taylor expansion around the Fermi energy E_F) can be applied to yield:

$$\alpha = \frac{\pi^2}{3} \frac{k_B}{e} k_B T \left\{ \frac{d[\ln(\sigma(E))]}{dE} \right\}_{E=E_F}. \quad (9)$$

When $\sigma(E)$ is attributed to band conduction (as opposed to hopping or strong localization) as is the case for thermoelectric materials, its relation to the density of states (DOS) $g(E)$, charge carrier concentration $n = \int g(E)f(E, E_F, T)dE$, and mobility μ is such that (Eq. 9) becomes:

$$\alpha = \frac{\pi^2}{3} \frac{k_B}{e} k_B T \left\{ \frac{g(E)}{n(E)} + \frac{1}{\mu(E)} \frac{d\mu(E)}{dE} \right\}_{E=E_F}. \quad (10)$$

Therefore, when $g(E)$ has a sharp maximum at $E = E_F$, the first factor in the thermopower is maximized, since $n(E)$ is essentially the area under the $g(E)$ curve up to $E = E_F$. When this formula is further worked into the ZT , classical thermoelectric theory [4] shows that the ZT is maximized in solids that have a high density of states effective mass. The traditional way to achieve this is to make use of solids where there the Fermi surface consists of several pieces, combining either a single heavy band or several separate ones.

More recent developments involved engineering the DOS energy dependence $g(E)$ in the

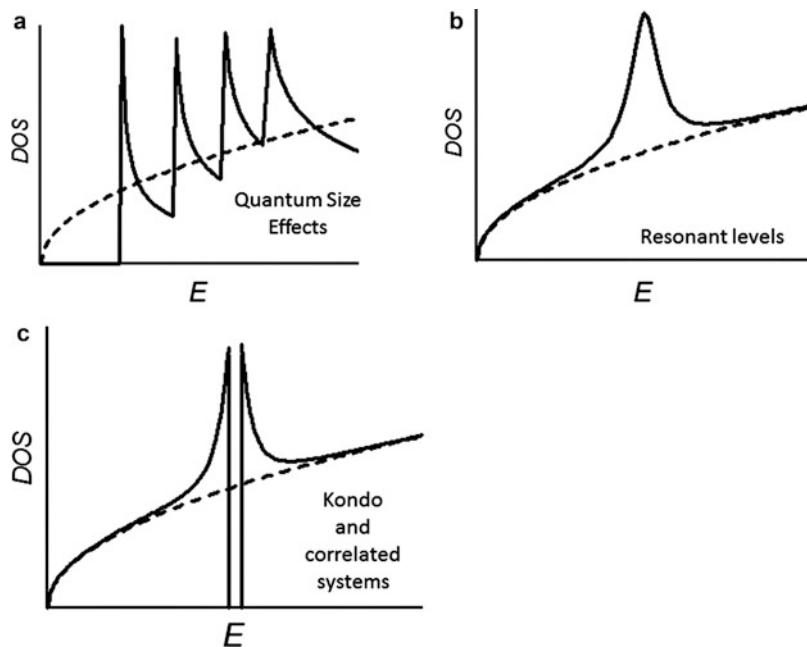
shapes shown in Fig. 7. In that figure, the DOS is shown for a conventional energy band as a dashed line. The techniques used to obtain sharp maxima in the DOS are illustrated frames (a)–(d) and are described in the following paragraph. Mahan and Sofo [35] point out one very important limitation: the effect of these approaches on ZT is strongly degraded when the background DOS (the dashed line) constitutes an important fraction of the DOS at E_F .

Size-Quantization

Semiconductor quantum wires are wires where the physical size of the structure is narrower than wavelength of the electrons (Fig. 7a). In such wires, electron motion is possible only in the long direction of the wire and quantized energy levels form in the transverse direction [37]. This was originally conceptualized by Hicks and Dresselhaus [34]. It has been implemented experimentally mainly in Bi [38–40] and Sb [41] nanowires. Strong enhancements of the thermopower were observed [42], but there are limitations on how narrow one can make these nanowires, as localization effects set in at around 10 nm diameters. In fact, in metals like Zn, where size-quantization effects require much narrower structures than in Bi, localization dominates [43]. Arrays of quantum wires in templates are difficult to use in thermoelectric materials, because the templates constitute a thermal short. Precipitating nanoparticles small enough that size-quantization effects inside them produce high thermopowers has also been tried using techniques similar to those described above. As a result, the thermopower of PbTe was shown to be enhanced [44].

Resonant Levels

Conventional doping of semiconductors consists of adding donor impurities or acceptor impurities to a host lattice (Fig. 7b). In a donor impurity, the impurity has one more electron on its outer shell than the atoms that constitute the host solid, such as P in Si. In an acceptor impurity, the impurity has one less electron on its outer shell than the atoms that constitute the host solid, such as Ga in Si. Consider only n-type doping of the conduction



Solid-State Heat Convertors, Fig. 7 Energy dependence of the electronic density of states (DOS) for several engineered band structures, designed to create favorable thermoelectric properties. The DOS of a conventional bulk semiconductor is shown as the dashed line. Frame (a) shows the DOS in quantum wires, where size-quantization squeezes the electron wave function into a

band for simplicity since the case of p-type doping is exactly symmetric. At very low temperature, the excess electron of the donor impurity stays localized on the donor, but it can be thermally excited and then become a mobile conduction electron. Translating this in the band model, one states that the excess electron has an energy state in the band gap of the semiconductor at an energy E_D below the conduction band edge (CB). The excess electron is localized on the band gap at low temperature, but it can be thermally activated from that band gap at temperatures T such that $k_B T > E_D$. The CB itself is determined by the properties of the host semiconductor, and the presence of the donor or acceptor does not modify it (the rigid band model, the dashed line in Fig. 3b). Some particular impurities in specific host semiconductors have electronic energy levels for which $E_D < 0$. The impurity level falls inside the conduction band. Its energy coincides with the energy of an extended state of the host. The two

one-dimensional shape. Frame (b) corresponds to the DOS in a system where a resonant impurity interacts with the band structure of the conventional semiconductor. Frame (c) is similar to frame (b), but now the interacting impurities add magnetic interactions to the system, and the peak splits in two peaks in which the electrons have opposite spin polarization

resonate to build up two extended states of slightly different energies; these in turn will have the same energies as other extended states with whom they will resonate in turn, and so on. Consequently, the resonant impurity [45] catalyzes the formation of an excess DOS (Fig. 7b). This excess DOS has, in fact, only very little of the character of the electronic levels of the impurity and dominantly consists of energy levels of the host solid, such that those states conduct electricity only slightly less than the original semiconductor. Nevertheless, because of their shape (Fig. 7b), and because of Eq. 10, they increase the thermopower, and thus the ZT , of the semiconductor (Tl in PbTe [46], Sn in Bi_2Te_3 [47]) in which they are observed.

Kondo and Correlated Systems

Figure 7c gives a very schematic representation of a concentrated Kondo/hybridized system [45].

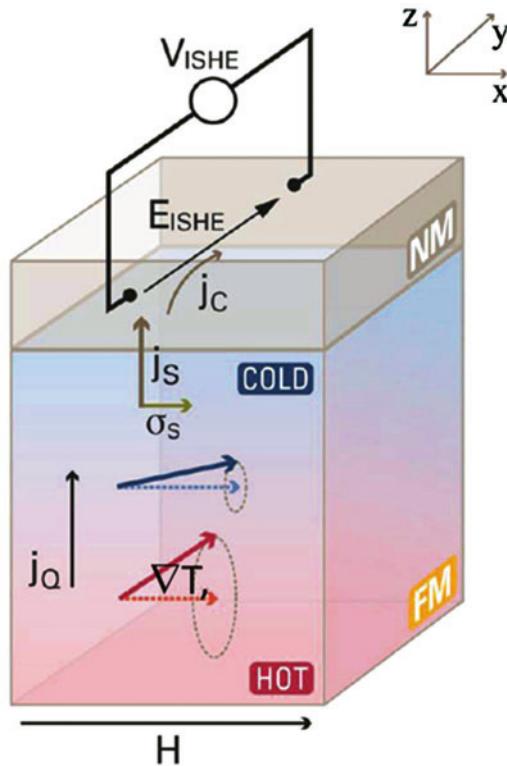
The boost in thermopower obtained from Kondo physics is an effect of the scattering of electrons (the second term in Eq. 10), which becomes indistinguishable from the DOS effect, and it is related to the spin polarization of that scattering – a link into the next chapter. The dilute Kondo effect (DK) was first observed in dilute alloys of metals where the dominant character of the bands correspond to s-orbitals (Cu, Ag, or Au) or p-orbitals to which small amounts transition metal impurities (Fe, Cr, Mn...) are added. The thermopower, which is proportional to T in metals, is greatly enhanced following a Gaussian peak in $\log(T)$ centered on a “Kondo temperature” T_K . T_K can reach 10^3 K [48], but the amplitude of the peak in S rarely exceeds a few tens of $\mu\text{V K}^{-1}$ in practice [49]. The resistivity, normally a monotonically decreasing function with decreasing T , has a minimum at T_K . Kondo explained [50] the effect on the basis of the fact that conduction electrons undergo a change of spin when they scatter on magnetic impurities, with the details rooted in the calculation of the scattering probabilities. Blatt et al. [51] represent the DOS in DK alloys as similar to that of the resonant levels (Fig. 7b), but with a magnetic splitting of the excesses in DOS into a spin-up and a spin-down bump.

When there are sufficient quantities of them, the magnetic atoms become a major constituent in the solid, such that their electronic levels contribute majorly to the band structure. When this occurs, the nearly dispersionless energy band that originates from the d - or f -levels of a rare-earth or transition metal constituent intersects a band with sp character in a semiconductor or metal. As a result, those two energy bands can hybridize [52]. A hybridization gap then opens in the DOS as depicted in Fig. 7c, and the local DOS near the gap is strongly distorted. Classical Kondo insulators are CeNiSn, CePd₃, and Ce₃Bi₄Pt₄ [53]. Many decades of research in this field have shown that hybridization gaps are not easy to use in the design of thermoelectric materials, particularly when they involve f -levels. Yet the CePd₂Pt [54] and YbAl₂ [55] systems hold some record ZT values in cryogenic cooling.

Spin

Very recently, spin has been added as a parameter to circumvent the difficulty inherent in thermoelectrics research, that is, the counter-indicated nature of the design parameters σ , κ , and α . Kondo systems rely on spin-dependent scattering, but the potential of thermal spin transport is still being investigated and the present understanding and future avenues of research are reviewed in this section. For the last 6 years, the exploration of spin caloritronic [56] effects has gained larger attention as a way of developing fundamentally new means for engineering thermal effects in materials, as well as providing methods to thermally generate pure spin fluxes. The first discovery in that field was the spin-Seebeck effect, found to exist in permalloy at Tohoku University in 2008 [57]. In 2010, the group at the Ohio State University discovered [58] the effect on a magnetic semiconductor, GaMnAs, and the Tohoku group found it on a ferromagnetic insulator and yttrium-iron garnet (YIG) at the same time [59]. Up to that time, the effect was measured in a geometry now called the transverse spin-Seebeck effect (TSSE). The Tohoku group discovered a simpler geometry on which the effect could be measured, the longitudinal Spin-Seebeck effect (LSSE). Because it can be confused with a conventional Nernst effect, the LSSE can be isolated only in electrically insulating ferromagnets [60]. Much larger TSSE effects were then found in InSb [61] using spin polarization of electrons by external magnetic fields rather than via ferromagnetism. Here, the giant spin-Seebeck effect can reach values of up to 8 mV/K, rivaling and even surpassing charge-based thermoelectrics, but only at cryogenic temperatures and high magnetic fields. The operating mechanism of the LSSE is now well established [62] and outlined next.

Figure 8 shows a diagram of an LSSE measurement using a YIG/Pt structure [56]. A 7 nm-thick Pt film is evaporated on the top surface of a crystal of YIG. A temperature gradient applied to that YIG from top to bottom generates a heat flux $j_Q = j_{QP} + j_{QM}$ that consists of heat carried by phonons (j_{QP}) and by spin waves or magnons (j_{QM}). The exact distribution between the phonon



Solid-State Heat Convertors, Fig. 8 Illustration of the longitudinal spin Seebeck effect (LSSE) geometry, a direct thermal spin injection from a magnon system in a ferromagnet (FM) into a normal metal (NM). The spin current polarizes the conduction electrons in the NM where, in the presence of spin-orbit interactions, it generates an electric field by the inverse spin-Hall effect (ISHE). *Energy Environ. Sci.*, 7 885 (2014) – Reproduced by permission of The Royal Society of Chemistry

and magnon conductivity was determined recently [63]. The magnon heat flux j_{QM} is directly related to the spin flux j_S carried by the magnons [56] $j_S = (\hbar j_{QM} / k_B T)$. This spin flux crosses the YIG/Pt interface at the top and spin-polarizes the conduction electrons in the Pt over a very small distance near the interface (~ 7 nm, the spin-diffusion length in Pt). Indeed, spin-flip transitions in the Pt quickly re-equilibrate the distribution of the electrons in the Pt so as to restore the same density of spin-up and spin-down electrons after one spin-diffusion length. The Pt film is thin enough that the electrons are spin-polarized through most of its thickness. Spin-polarized electrons in Pt are subject to the inverse spin-Hall

effect (ISHE) [56]: due to skew scattering and other intrinsic effects, in solids made from heavy atoms, electrons scatter at different angles depending on their spin polarization. If there are more spin-up than spin-down electrons, the excess electrons accumulate on one side of the sample, creating an electric field proportional to the flux of spin-polarized electrons, which is equal to the spin flux. This is the inverse spin-Hall field $E_{ISHE} = D(\mathbf{j}_S \times \boldsymbol{\sigma})$, where, in vector notation, \mathbf{j}_S is the vector of spin flux (which, as we have seen above is thermally driven) and $\boldsymbol{\sigma}$ is the spin polarization vector (parallel to the applied magnetic field \mathbf{H}). Figure 8 shows the orientations of all these vectors. The end result is that a voltage appears across the Pt when a temperature gradient is applied to the ferromagnet, just like in a conventional thermoelectric or, more precisely, like in a Nernst effect measurement (the electric field induced by the Nernst effect, like E_{ISHE} , is perpendicular to the temperature gradient). While it has been shown [61] in InSb that the field to temperature gradient ratio (the spin-Seebeck coefficient) can be as large as for classical thermopower effects, calculations [64] of the ZT of the YIG/Pt system is very poor (10^{-3}). The causes are reviewed here, as well as avenues of ongoing research.

The two main steps for the energy conversion is the transformation of the heat flux into a spin flux, and then the spin flux into a charge flux and voltage. The first step is similar to the case of classical thermoelectrics: the heat carried by electrons or magnons is equivalent to a charge or spin flux, both of which are useful, but the heat carried by phonons (the lattice) is a pure loss. Next, there are three loss mechanisms specific to the LSSE, and further research could prove that all three may be circumvented. First, the spin flux must cross the ferromagnet/metal interface, which is characterized by a parameter called the spin-mixing conductance ($g_{\uparrow\downarrow}$). This parameter is very well characterized [65]. Its physics is based on the interactions between the conduction electrons in the Pt and the core d-electrons on the ferromagnetic atoms in the YIG. It is quite akin to the mechanisms that limit the electrical resistivity of transition metals like Fe, where conduction

electrons interact with core d-electrons. The fact that this mechanism leads to a resistivity that is 10–100 times higher in transition metals than in copper proves that this is very efficient. While $g_{\uparrow\downarrow}$ depends on the quality of the interface, generally it is not a major cause of loss of spin polarization. The second loss is the conversion of spin flux into a voltage, i.e., the efficiency of the inverse spin-Hall effect. A review of this effect is given by Hoffmann [66]. From an engineering perspective, the conversion efficiency – included in the parameter D in the formula above – is at most 1–3 %, except in some narrow-gap semiconductors (InSb) which are not well studied yet. This loss is specific to LSSE. The third, and by far the worst, loss mechanism in the LSSE is the fact that only a very small fraction of the volume of the ferromagnet contributes to the LSSE. This was proven by Kehlberger et al. [67] who measured the LSSE as in Fig. 8, but using films of YIG grown on a nonmagnetic substrate (a gadolinium gallium garnet, GGG). They plotted the spin-Seebeck coefficient as a function of YIG film thickness and showed that the intensity of the effect increased with film thickness up to about 150 nm and saturated thereafter. This proves that only the spin flux generated in the 150 nm layer under the Pt gets converted into electrical energy. Because the temperature gradient is applied across the entire film plus substrate thickness (0.5 mm), only 0.03 % of that gradient does any useful work. This is the worst energy loss mechanism, but also the easiest to address: one could make multilayers of 100 nm YIG/7 nm Pt [68] – leaving the losses in the spin-Hall conversion as dominant loss mechanism. Such work is ongoing.

Finally, there is a connection between spin caloritronics and conventional thermoelectrics as pertains to metals: magnon drag. Usual metals have a thermopower limited to $|S| < 10 \mu\text{V K}^{-1}$. In reality, metals used in thermocouples have values as high as $\pm 50 \mu\text{V K}^{-1}$ or even higher; but even those values are not sufficient to give a good ZT . One thermocouple material is iron, with $\alpha > 15 \text{ V K}^{-1}$ at 300 K. As far back as the 1960s, this unusual behavior was speculatively interpreted as arising from a spin-based mechanisms called magnon drag [69]. We now

understand [70] that magnon drag is in fact a form of self-spin-Seebeck effect, where the electrons that give rise to the voltage are not in a layer adjacent to the ferromagnet, but in the ferromagnet itself. Such an effect was observed on metallic glasses [71].

There is an extensive theoretical literature on magnon drag in metals, but the basic hydrodynamic theory [72] is presented here and similar to the McDonald theory [73] for phonon drag. Consider both magnons in ferromagnets and electrons as quasi-particles with properties similar to ideal gasses, including the fact that they have parabolic dispersion relations between energy and momentum. Note that the case of phonons and of magnons in antiferromagnets, which have linear dispersions, can be treated very similarly, but differ by a numerical factor of two. In all drag situations, the “dragging” quasi-particle, here the magnon, is assumed to undergo collisions that are predominantly with the “dragged” particles, here the electrons. This means that we consider the situation where magnon-electron interactions, at frequency τ_{ME}^{-1} , constitute the dominant magnon scattering mechanism, i.e., $\tau_{ME}^{-1} \approx \tau_M^{-1}$ (the total magnon collision frequency being τ_M^{-1}). Consider the magnons as an isotropic ideal gas, with an internal energy U : the magnons will exert a pressure on the electrons with which they collide given by $p = 2/3U$. Now, apply a temperature gradient dT/dx along a specific direction x : this will give rise to a pressure gradient in the magnon population. The pressure gradient will give rise to a force F_x per unit volume, selectively applied to the electron population, that, since the problem is one-dimensional, is given by:

$$F_x = -\frac{dp}{dx} = -\frac{2}{3} \frac{dU}{dx} = -\frac{2}{3} \frac{dU}{dT} \frac{dT}{dx}. \quad (11)$$

This is the force that impels a unidirectional momentum onto the electron population, so that their momentum along the direction x , i.e., $\hbar k_x$, is modified. This will give rise to an instantaneous magnon-drag thermoelectrical current in the sample proportional to $-dT/dx$. When no current is permitted to flow, it will result in an accumulation of electrons and the creation of an electric field E_x

that will act on the electrons to generate a force such that the force neE_x . The balance of forces means that:

$$\begin{aligned} F_x + neE_x &= 0 \\ neE_x &= \frac{2}{3} \frac{dU}{dT} \frac{dT}{dx} = \frac{2}{3} C_M \frac{dT}{dx}. \end{aligned} \quad (12)$$

where C_M is the magnon-specific heat. Given the definition of the thermopower α as $E_x/(dT/dx)$, it results that its magnon-drag component α_{MED} is then simply:

$$\alpha_{MED} = \frac{2C_M}{3ne}. \quad (13)$$

Reconsider now the situation where there are other mechanisms that scatter magnons besides interactions with electrons, i.e., where $\tau_{ME}^{-1} < \tau_M^{-1}$. The other scattering mechanisms will absorb a fraction of the magnon dragging force that would have been impelling the electrons in the above discourse, reducing it by a fraction $\tau_{ME}^{-1}/\tau_M^{-1}$. The same fraction then is reflected in the magnon-drag thermopower, which becomes:

$$\alpha_{MED} = \frac{2C_M}{3ne} \frac{\tau_M}{\tau_{ME}}. \quad (14)$$

In reality, one has to consider the above argument for each specific mode and energy of the entire magnon and electron populations, so that the exact solutions for α_{MED} involve integrals over the whole Brillouin zone of both electron and magnon dispersions, and this is what is done in the specialized literature. Nevertheless, one can extract important physical conclusions about α_{MED} from just Eq. 14. First, drag becomes more important as the carrier concentration is reduced; for the case of phonon drag, values exceeding 10 mV/K are easily observed in Si and Ge. Second, at low temperature and in large, perfect crystals where the condition $\tau_{ME}^{-1} \approx \tau_M^{-1}$ holds, magnon-drag is expected to mimic the temperature dependence of $C_M \propto T^{3/2}$, just like phonon drag follows the Debye lattice-specific heat, a T^3 law. Third, at a sufficiently high temperature where $\tau_{ME}^{-1} < \tau_M^{-1}$, α_{MED} will follow the

temperature dependence of $C_M \times \tau_M$, because there is not much temperature dependence to magnon-electron interactions, which are generally due to s-d interactions. It is on this point that magnon drag differs strongly from phonon drag: in the case of phonons, the Debye-specific heat saturates at the Debye temperature Θ_D . Further, in the very pure crystals where phonon drag is observed, anharmonic scattering is the dominant mechanism limiting the lattice thermal conductivity for all temperatures $T > \Theta_D/3$ to $\Theta_D/10$. Phonon-drag therefore is a low-temperature phenomenon in most solids with $\Theta_D < 1000$ K. Not so for magnon drag, because the magnon density keeps increasing up to the Curie temperature, which can be quite high in ferromagnetic metals, and also because s-d scattering remains strong at all temperatures. Magnon-drag or other spin-mediated thermopowers in ferromagnetic metals easily extend to and exceed room temperature.

Conclusions

In summary, the Hicks and Dresselhaus papers [33, 34] gave a renewed impetus to research on thermoelectric solid-state energy conversion, which has great potential to recover heat otherwise wasted into the environment and convert it to useful energy or to provide thermodynamic topping cycles that increase the thermal efficiencies. The last 20 years of research have resulted in an increase of the thermoelectric figure of merit from that value it was at in the 1970s $ZT \approx 1$, to the value obtained on SnSe this summer, $ZT \approx 2.6$ [28]. This ZT reflects the increase in thermal efficiency of future devices that could be built out of the new materials, although the development of thermoelectric devices has not kept pace with the scientific discoveries yet [12]. The role of spin in thermoelectrics and spin caloritronics is a new avenue, largely unexplored, that promises more progress. Speculatively, perhaps future emphasis could be placed on materials that are more earth abundant than the elements currently used, and also easier to produce, shape, and contact than semiconductors. A good place to start would be to develop high- ZT metals.

Acknowledgments The author acknowledges help with the manuscript editing and figures from Ms. Renee Ripley and support from the NSF MRSEC program, Grant No. DMR 1420451.

Cross-References

- ▶ Heat Capacity
- ▶ Heat Conduction
- ▶ Heat Conductivity
- ▶ Heat Transfer in Semiconductor Nanostructures
- ▶ Nanostructures for Energy
- ▶ Spintronic Devices
- ▶ Thermal Conductance
- ▶ Thermal Conductivity and Phonon Transport
- ▶ Thermal Resistance
- ▶ Thermal Resistivity
- ▶ Thermal Transport

References

1. Telkes, M.: The efficiency of thermoelectric generators. *I. J. Appl. Phys.* **18**(12), 1116–1127 (1947)
2. Telkes, M.: Power output of thermoelectric generators. *J. Appl. Phys.* **25**(8), 1058–1059 (1954)
3. Ioffe, A.F.: *Semiconductor Thermoelements and Thermoelectric Cooling*. Inforsearch, London (1957)
4. Goldsmid, H.J.: *Thermoelectric Refrigeration*. Plenum Press, New York (1964)
5. Seebeck, T.J.: Magnetische Polarisation der Metalle und Erze durch Temperatur-Differenz. In: Abhandlungen der Preussischen Akad, Wissenschaften, pp. 265–373 (1822–1823). Reprinted W. Engelmann, Leipzig (1895)
6. Thomson, W.: On the dynamical theory of heat. Part V. Thermo-electric currents. *Trans. R. Soc. Edinb.* **21**, 123–171 (1857). <https://archive.org/stream/transactionsofro21royal#page/n3/mode/2up>
7. Lord Rayleigh, F.R.S.: On the thermodynamic efficiency of the thermopile. *Philos. Mag.* **20**, 361–363 (1885)
8. Altenkirch, E.: Über den Nutzeffekt der Thermosäule. *Phys. Ztg.* **10**(16), 560–568 (1909)
9. Callen, H.B.: *Thermodynamics and an Introduction to Thermostatistics*. Wiley, New York (1960)
10. Roberts, R.B.: Absolute scales for thermoelectricity. *Measurement* **4**(3), 101–103 (1986). doi:10.1016/0263-2241(86)90016-3; Roberts, R.B.: Absolute scale of thermoelectricity. *Philos. Mag.* **36**, 91 (1977); Roberts, R.B.: Absolute scale of thermoelectricity II. *Philos. Mag. B* **43**, 1123 (1981); Roberts, R.B., Righini, F., Compton, R.C.: Absolute scale of thermoelectricity III. *Philos. Mag. B* **52**, 1147 (1985)
11. Vining, C.B.: An inconvenient truth about thermoelectrics. *Nat. Mater.* **8**, 83–85 (2009)
12. Heremans, J.P., Dresselhaus, M.S., Bell, L., Morelli, D.T.: When thermoelectrics reached the nanoscale. *Nat. Nanotechnol.* **8**, 471–473 (2013)
13. Ioffe, A.F.: *Physics of Semiconductors*. Academic, New York (1960)
14. Berman, R.: *Thermal Conduction in Solids*. Clarendon, Oxford (1976)
15. Morelli, D.T., Heremans, J.P., Slack, G.A.: Estimation of the isotope effect on the lattice thermal conductivity of group IV and group III-V semiconductors. *Phys. Rev. B* **66**, 195304 (2002)
16. Poudel, B., Hao, Q., Ma, Y., Lan, Y., Minnich, A., Yu, B., Yan, X., Wang, D., Muto, A., Vashaee, D., Chen, X., Liu, J., Dresselhaus, M.S., Chen, G., Ren, Z.: High-thermoelectric performance of nanostructured bismuth antimony telluride bulk alloys. *Science* **320**, 634–638 (2008)
17. Lindsay, L., Broido, D.A., Reinecke, T.L.: Ab initio thermal transport in compound semiconductors. *Phys. Rev. B* **87**, 165201 (2013)
18. Morelli, D.T., Meissner, G.P.: Low temperature properties of the filled skutterudite CeFe₄Sb₁₂. *J. Appl. Phys.* **77**, 3777 (1995)
19. Meissner, G.P., Morelli, D.T., Hu, S., Yang, J., Uher, C.: Structure and lattice thermal conductivity of fractionally filled skutterudites: solid solutions of fully filled and unfilled end members. *Phys. Rev. Lett.* **80**, 3551 (1998)
20. Sales, B.C., Mandrus, D., Williams, R.K.: Filled skutterudite antimonides: a new class of thermoelectric materials. *Science* **272**, 1325–1328 (1996)
21. Shi, X., Yang, J., Salvador, J.R., Chi, M.F., Cho, J.Y., Wang, H., Bai, S.Q., Yang, J.H., Zhang, W.Q., Chen, L.D.: Multiple-filled skutterudites: high thermoelectric figure of merit through separately optimizing electrical and thermal transports. *J. Am. Chem. Soc.* **133**, 7837 (2011)
22. Nielsen, M.D., Ozolins, V., Heremans, J.P.: Lone pair electrons minimize lattice thermal conductivity. *Energy Environ. Sci.* **6**(2), 570–578 (2013)
23. Jovovic, V., Heremans, J.P.: Doping effects on the thermoelectric properties of AgSbTe₂. *J. Electron. Mater.* **38**, 1504–1509 (2009)
24. Jovovic, V., Heremans, J.P.: Energy band gap and valence band structure of AgSbTe₂. *Phys. Rev. B* **77**, 245204 (2008)
25. Morelli, D.T., Jovovic, V., Heremans, J.P.: Intrinsically minimal thermal conductivity in cubic I-V-VI₂ semiconductors. *Phys. Rev. Lett.* **101**, 035901 (2008)
26. Lu, X., Morelli, D.T., Xia, Y., Zhou, F., Ozolins, V., Chi, H., Uher, C.: High performance thermoelectricity in earth-abundant compounds based on natural mineral tetrahedrites. *Adv. Energy Mater.* **3**, 342–348 (2012)
27. Delaire, O., Ma, J., Marty, K., May, A.F., McGuire, M.A., Du, M.-H., Singh, D.J., Podlesnyak, A., Ehlers, G., Lumsden, M.D., Sales, B.C.: Giant anharmonic

- phonon scattering in PbTe. *Nat. Mater.* **10**, 614–619 (2011)
28. Zhao, L.-D., Lo, S.-H., Zhang, Y., Sun, H., Tan, G., Uher, C., Wolverton, C., Dravid, V.P., Kanatzidis, M.G.: Ultralow thermal conductivity and high thermoelectric figure of merit in SnSe crystals. *Nature* **508**, 373–377 (2014)
 29. Heremans, J.P.: Thermoelectricity: the ugly duckling. *Nature* **508**, 327–328 (2014)
 30. Heremans, J.P., Thrush, C.M., Morelli, D.T.: Thermopower enhancement in PbTe with Pb precipitates. *J. Appl. Phys.* **98**, 063703 (2005)
 31. Girard, S.N., He, J., Zhou, X.Y., Shoemaker, D., Jaworski, C.M., Uher, C., Dravid, V.P., Heremans, J.P., Kanatzidis, M.G.: High performance Na-doped PbTe-PbS thermoelectric materials: electronic density of states modification and shape-controlled nanostructures. *J. Am. Chem. Soc.* **133**, 16588–16597 (2011)
 32. Biswas, K., He, J.Q., Blum, I.D., Chun, I.W., Hogan, T.P., Seidman, D.N., Dravid, V.P., Kanatzidis, M.G.: High-performance bulk thermoelectrics with all-scale hierarchical architectures. *Nature* **490**, 414–418 (2012)
 33. Hicks, L.D., Dresselhaus, M.S.: Effect of quantum-well structures on the thermoelectric figure of merit. *Phys. Rev. B* **47**, 12727–12731 (1993)
 34. Hicks, L.D., Dresselhaus, M.S.: Thermoelectric figure of merit of a one-dimensional conductor. *Phys. Rev. B* **47**, 16631–16634 (1993)
 35. Mahan, G.D., Sofo, J.O.: The best thermoelectric. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 7436–7439 (1996)
 36. Cutler, M., Mott, N.F.: Observation of Anderson localization in an electron gas. *Phys. Rev.* **181**, 1336 (1969)
 37. Heremans, J.P.: Low-dimensional thermoelectricity. *Acta Phys. Polon.* **108**, 609–634 (2005)
 38. Murata, M., Nakamura, D., Hasegawa, Y., Komine, T., Taguchi, T., Nakamura, S., Jovovic, V., Heremans, J.P.: Thermoelectric properties of bismuth nanowires in a quartz template. *Appl. Phys. Lett.* **94**, 192104 (2009)
 39. Heremans, J., Thrush, C.M.: Thermoelectric power of bismuth nanowires. *Phys. Rev. B* **59**, 12579 (1999)
 40. Murata, M., Nakamura, D., Hasegawa, Y., Komine, T., Taguchi, T., Nakamura, S., Jaworski, C.M., Jovovic, V., Heremans, J.P.: Mean free path limitation of thermoelectric properties of bismuth nanowire. *J. Appl. Phys.* **105**, 113706 (2009)
 41. Heremans, J., Thrush, C.M., Lin, Y.-M., Cronin, S.B., Dresselhaus, M.S.: Transport properties of antimony nanowires. *Phys. Rev. B* **63**, 085406 (2001)
 42. Heremans, J.P., Thrush, C.M., Morelli, D.T., Wu, M.-C.: Thermoelectric power of bismuth nanocomposites. *Phys. Rev. Lett.* **88**, 216801 (2002)
 43. Heremans, J.P., Thrush, C.M., Morelli, D.T., Wu, M.-C.: Resistance, magnetoresistance and thermopower of zinc nanowire composites. *Phys. Rev. Lett.* **91**, 076804 (2003)
 44. Heremans, J.P., Thrush, C.M., Morelli, D.T.: Thermopower enhancement in lead telluride nanostructures. *Phys. Rev. B* **70**, 115334 (2004)
 45. Heremans, J.P., Wiendlocha, B., Chamoire, A.M.: Resonant levels in bulk thermoelectric semiconductors. *Energy Environ. Sci.* **5**, 5510–5530 (2012)
 46. Heremans, J.P., Jovovic, V., Toberer, E.S., Saramat, A., Kurosaki, K., Charoenphakdee, A., Yamanaka, S., Snyder, G.J.: Enhancement of thermoelectric efficiency in PbTe by distortion of the electronic density of states. *Science* **321**, 554–558 (2008)
 47. Jaworski, C.M., Kulbachinskii, V.A., Heremans, J.P.: Tin forms a resonant level in Bi₂Te₃ that enhances the room temperature thermoelectric power. *Phys. Rev. B* **80**, 233201 (2009)
 48. Daybell, M.D., Steyert, W.A.: Localized magnetic impurity states in metals: some experimental relationships. *Rev. Mod. Phys.* **40**, 380 (1968)
 49. Heeger, A.J.: Localized moments and nonmoments in metals. In: Seitz, F., Turnbull, D., Ehrenreich, H. (eds.) *Solid State Physics*, vol. 23, pp. 284–407. Academic, New York (1969)
 50. Kondo, J.: Resistance minimum in dilute magnetic alloys. *Prog. Theor. Phys.* **34**, 372 (1965)
 51. Blatt, F.J., Schroeder, P.A., Foiles, C.L., Greig, D.: *Thermoelectric Power of Metals*. Plenum Press, New York (1976)
 52. Mahan, G.D.: Good thermoelectrics. In: Ehrenreich, H., Spaepen, F. (eds.) *Solid State Physics*, vol. 51, pp. 81–152. Academic, New York (1997)
 53. Fisk, Z., Sarrao, J.L., Thompson, J.D.: Heavy fermions. *Curr. Opin. Solid State Mater. Sci.* **1**, 42 (1996), Ce₃Bi₄Pt₃ and CePd₃ 42 (1996)
 54. Boona, S.R., Morelli, D.T.: Enhanced thermoelectric properties of CePd_{3-x}Pt_x. *Appl. Phys. Lett.* **101**, 101909 (2012)
 55. Lehr, G.J., Morelli, D.T., Jin, H., Heremans, J.P.: Enhanced thermoelectric power factor in Yb_{1-x} Sc_x Al₂ alloys using chemical pressure tuning of the Yb valence. *J. Appl. Phys.* **114**, 223712 (2013)
 56. Boona, S.R., Myers, R.C., Heremans, J.P.: Spin caloritronics. *Energy Environ. Sci.* **7**, 885–910 (2014). doi:10.1039/C3EE43299H
 57. Uchida, K., Takahashi, S., Harii, K., Ieda, J., Koshibae, W., Ando, K., Maekawa, S., Saitoh, E.: Observation of the spin Seebeck effect. *Nature* **455**, 778–781 (2008)
 58. Jaworski, C.M., Yang, J., Mack, S., Awschalom, D.D., Heremans, J.P., Myers, R.C.: Observation of the spin-Seebeck effect in a ferromagnetic semiconductor. *Nat. Mater.* **9**, 898–903 (2010)
 59. Uchida, K., Xiao, J., Adachi, H., Ohe, J., Takahashi, S., Ieda, J., Ota, T., Kajiwara, Y., Umezawa, H., Kawai, H., Bauer, G.E.W., Maekawa, S., Saitoh, E.: Spin Seebeck Insulator. *Nat. Mater.* **9**, 894–897 (2010). doi:10.1038/NMAT2856
 60. Uchida, K.-i., Adachi, H., Ota, T., Nakayama, H., Maekawa, S., Saitoh, E.: Observation of longitudinal

- spin-Seebeck effect in magnetic insulators. *Appl. Phys. Lett.* **97**, 172505 (2010)
61. Jaworski, C.M., Myers, R.C., Johnston-Halperin, E., Heremans, J.P.: Giant spin Seebeck effect in a non-magnetic material. *Nature* **487**, 210–213 (2012)
 62. Hoffman, S., Upadhyaya, P., Tserkovnyak, Y.: Landau-Lifshitz theory of the longitudinal spin Seebeck effect. *Phys. Rev. B* **88**, 064408 (2013)
 63. Boona, S.R., Heremans, J.P.: Magnon thermal mean free path in yttrium iron garnets. *Phys. Rev. B* **90**, 064421 (2014). doi:10.1103/PhysRevB.90.064421
 64. Kovalev, A.A., Tserkovnyak, Y.: Magnetocaloritronic nanomachines. *Solid State Commun.* **150**, 500–504 (2010)
 65. Weiler, M., Althammer, M., Schreier, M., Lotze, J., Pernpeintner, M., Meyer, S., Huebl, H., Gross, R., Kamra, A., Xiao, J., Chen, Y.-T., Jiao, H.J., Bauer, G.E.W., Goennenwein, S.T.B.: Experimental test of the spin mixing interface conductivity concept. *Phys. Rev. Lett.* **111**, 176601 (2013)
 66. Hoffmann, A.: Spin Hall effects in metals. *IEEE Trans. Magn.* **49**, 5172–5193 (2013)
 67. Kehlberger, A., Ritzmann, U., Hinzke, D., Guo, E.-J., Cramer, J., Jakob, G., Onbasali, M. C., Kim, D. H., Ross, C. A., Jungfleisch, M. B., Hillebrands, B., Nowak, U., and Kläui, M.: Length scale of the spin Seebeck effect, *Phys. Rev. Lett.* **115**, 096602 (2015)
 68. Heremans, J., Jaworski, C.: Spin thermoelectric generator with multiple magnetic layers. Ohio State University Technology Commercialization Office, Invention disclosure T2012-251 (2012)
 69. Blatt, F.J., Flood, D.J., Rowe, V., Schroeder, P.A., Cox, J.E.: Magnon-drag thermopower in iron. *Phys. Rev. Lett.* **18**, 395 (1967)
 70. Lucassen, M.E., Wong, C.H., Duine, R.A., Tserkovnyak, Y.: Spin-transfer mechanism for magnon-drag thermopower. *Appl. Phys. Lett.* **99**, 262506 (2011)
 71. Jin, H., Yang, Z., Myers, R.C., Heremans, J.P.: Spin-Seebeck like signal in ferromagnetic bulk metallic glass without platinum contacts. *Solid State Commun.* **198**(Special Issue on Spin Mechanics), 40–44 (2014)
 72. Watzman, S.J., Duine, R.A., Tserkovnyak, Y., Jin, H., Prakash, A., Zheng, Y., and Heremans, J.P.: Magnon-drag thermopower and Nernst coefficient in Fe and Co, arXiv 1603.03736 (2016)
 73. MacDonald, D.C.K.: *Thermoelectricity: An Introduction to the Principles*. Wiley, New York (1962)

Sound Propagation in Fluids

- [Acoustic Nanoparticle Synthesis for Applications in Nanomedicine](#)

Spectromicroscopy

- [Selected Synchrotron Radiation Techniques](#)

Spectroscopic Techniques

- [Optical Techniques for Nanostructure Characterization](#)

Spectroscopy of Ancient Documents

M. Missori¹, O. Pulci¹ and A. Mosca Conte²

¹Institute for Complex Systems, National Research Council, Rome, Italy

²ETSF, and Department of Physics, University of Rome Tor Vergata, Rome, Italy

Synonyms

[Optical reflectance spectroscopy of ancient paper](#); [Optical spectroscopy for ancient paper diagnostic](#)

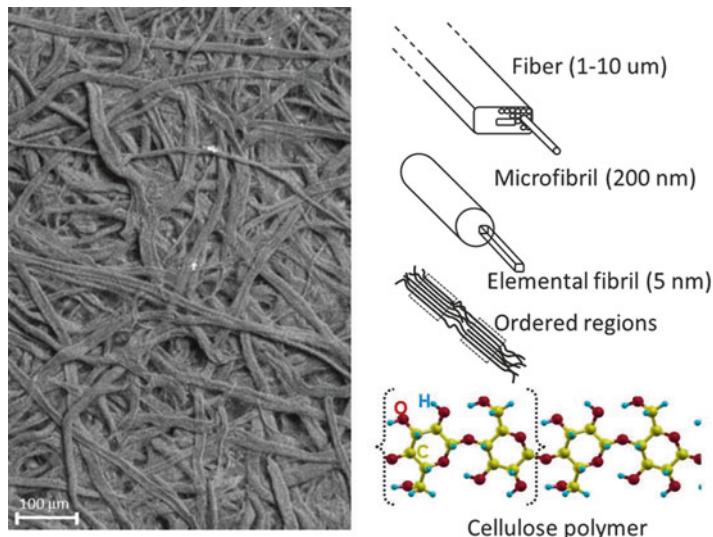
Definition

Optical spectroscopy of ancient paper is the study of the interaction between paper and ultraviolet, visible and near infrared radiation. This approach is aimed to nondestructive evaluation of the degradation of ancient artifacts.

Cellulose and Ancient Paper

Cellulose has had a unique role in civilization because it has been widely used for the acquisition, storage, and dissemination of human culture. In fact, one of the main cellulose derivatives is paper, which was invented in China in the second century A.D. and, independently, in Mesoamerica in the seventh century A.D. The technological

Spectroscopy of Ancient Documents, Fig. 1 Image of Whatman no. 1 paper sample, without prior preparation, obtained by using a Zeiss EVO40 variable pressure electron microscope. A random network of many thick fibers tangled together to form a complex structure can be observed (*left*). Schematic view of the supermolecular architecture of cellulose polymers within a fiber (*right*)



knowledge of making paper spread, in the Middle Ages, from the Far East to Arab World and Europe. As a result, for centuries, a growing number of books, documents, and artistic drawings have been accumulating in archives, libraries, and museums, all over the world [1].

The preservation of this cultural heritage is based on limiting the degradation of the paper materials. To this aim, optical spectroscopy allows an advanced knowledge of the nanoscale characteristics of paper materials and their degradation processes.

Paper is a complex multicomponent material consisting mainly of cellulose fibers. A paper sheet is obtained from a dilute suspension of cellulose fibers in water that are then drained through a sieve, pressed and dried, to obtain a network of randomly interwoven fibers. The paper composition varies depending on the production period and technology employed. In the Middle Ages in Europe, paper was made up of pure cellulose fibers (>90 % in weight) from cotton, linen, or hemp, usually obtained from rags with the addition of animal glue as a sizing agent [1]. Besides cellulose, most of modern paper also contains a relevant quantity of lignin and fillers.

Cellulose is the most abundant organic compound on Earth representing 40 % of the annual production of biomass [2]. It is a linear homopolymer composed of β -D-glucopyranose units

($C_6H_{10}O_5$)_n, which are linked together by β -(1,4)-glycosidic bonds up to form chains with n from about 100 to about 10,000 elements (see Fig. 1). The cellulose chains have a strong tendency to aggregate into highly ordered structural entities through an extended network of both intra- and intermolecular hydrogen bonds. As a consequence, a hierarchical arrangement of polymers is formed, from elemental fibrils through microfibrils up to fibers, whose diameter ranges approximately from 1 to 10 μm [3, 4].

Elemental fibrils, basic components of the cellulose supramolecular structure, include assembly of highly ordered (crystalline) domains and of disordered (amorphous-like) regions [2–4] (Fig. 1). Crystalline domains constitute from around 60 % to 70 %, of total cellulose material depending on its origin and history.

Aging of Cellulose and Paper

On macroscopic scale, paper degradation occurs by the weakening of mechanical properties of the sheets and by yellowing. At the nanoscale, cellulose degradation can be seen as the combination of two most important processes: acid hydrolysis of β -(1,4)-glycosidic bonds, and oxidation of the β -D-glucopyranose units with the subsequent development of various chemical products [3, 5].

The degradation extent in paper after a period of time depends on the environmental conditions to which it has been subjected [5]. While the mechanism of acidic hydrolysis is rather known and well settled in the literature, cellulose oxidation running through the radical mechanism initiated by active oxygen species is a complex process with many possible reactions still to be clarified [2–5].

Pristine cellulose does not absorb light up to about 200 nm (below 6 eV of photon energy) [5]. Paper yellowing is due to some oxidation products which are called chromophores. The yellow color seen in aged papers is mainly due to the fact that chromophores in paper absorb the higher energy band of visible light (corresponding to violet and blue) and largely scatter the yellow and red portion, thereby producing the characteristic yellow-brown hue [6].

Optical Spectroscopy of Inhomogeneous Materials

A medium is optically inhomogeneous when it is characterized by variations of the dielectric function ϵ over distances comparable to those of the wavelength of light impinging on the sample. This induces the scattering of light, namely, the deflection of photons from a straight path [7]. As a consequence, the usual laws related to optical properties of homogeneous materials are no more valid [8]. In general, scattering of photons is caused by, for example, irregularities in the propagation medium, presence of particles, or roughness in the interface between two media. In the presence of scattering, the reflectivity of materials is commonly referred to as diffuse reflectance [8]. Paper sheets being composed by rough-surfaced fibers and voids represent a striking example of materials whose optical properties are strongly governed by light-scattering effects.

In order to separate the scattering from the absorption, a number of theories have been developed [8]. Among those, some phenomenological models have been applied to describe paper optical properties [9]. However, the most successful approach is the Kubelka-Munk (KM) model [8],

an approach using two radiative fluxes to the general theory of radiation transfer in optically inhomogeneous substances (such as a sheet of paper). Using this model, it is possible to correlate reflectance measurements to the absorption coefficient of the elements that make up the inhomogeneous material.

KM model assumes that light propagates in just two opposing directions in an inhomogeneous material, with the flux variation at any point in the medium being linearly proportional to the two local opposing fluxes. The proportionality constants called scattering, S , and absorption, K , coefficients are assumed dependent on the intrinsic scattering and absorption properties of the medium. The radiative flux I which propagates downward ($-z$ direction) in a material represents the averaged intensity of all rays directed toward the lower hemisphere. Similarly, the flow J which propagates upward ($+z$ direction) is the average intensity of all rays directed toward the upper hemisphere. The behavior of J and I is described by the system of differential equations:

$$\begin{aligned} -\frac{dI}{dz} &= -(K + S)I + SJ \\ \frac{dJ}{dz} &= -(K + S)J + SI \end{aligned} \quad (1)$$

where z is the upward coordinate of the thickness of the material and S and K coefficients have both physical dimensions of the reciprocal of length. The system of Eq. 1 can be solved analytically and provides the so-called *Kubelka-Munk function*:

$$A_{KM} = \frac{K}{M} = \frac{(1 - R_\infty)^2}{2R_\infty} \quad (2)$$

where R_∞ is the measured reflectance, i.e., the ratio of the reflected flux to the incident flux at the sample surface, when the sample is sufficiently thick that no incident radiation passes through it.

K and S coefficients can also be calculated individually for thin samples from KM model. To this aim, two separate reflectance measurements obtained by laying the sheets on a white (R_w) backing, and then a black (R_b) backing must

be carried out. The reflectance values for the white (R_{wb}) and black (R_{bb}) backings must be measured separately.

Hence, the reflectance R_∞ that an infinitely thick layer of the same sample would have, can be calculated by using the results of the original KM model [8, 9], from R_w , R_b , R_{wb} , and R_{bb} :

$$R_\infty = a - \sqrt{a^2 - 1} \quad (3)$$

and

$$a = \frac{1}{2} \frac{(R_{wb} - R_{bb})(1 + R_w R_b) - (R_w - R_b)(1 + R_{wb} R_{bb})}{R_b R_{wb} - R_w R_{bb}} \quad (4)$$

When R_∞ is known, it is possible to derive the phenomenological scattering S and absorption K coefficients:

$$S = \frac{1}{t \left(\frac{1}{R_\infty} - R_\infty \right)} \ln \frac{(1 - R_b R_\infty)(R_\infty - R_{bb})}{(1 - R_\infty R_{bb})(R_\infty - R_b)} \quad (5)$$

and

$$K = \frac{S(1 - R_\infty)^2}{2R_\infty} \quad (6)$$

where t is the thickness of the inhomogeneous sample [8].

The KM theory has been extended by Yang and Miklavcic (YM) to take into account samples with stronger optical absorption [10]. Following this approach, the intrinsic absorption α and scattering s coefficients of paper, in the case of uniform distribution of the incident and reflected light (obtained, e.g., by means of an integrating sphere), can be recovered as

$$\begin{aligned} \alpha &= \frac{K}{2\mu} \\ s &= \frac{S}{\mu} \end{aligned} \quad (7)$$

The quantity $\mu = \mu(s, \alpha)$ is the scattering-induced path variation (SIPV) factor [10]. It describes the influence of light scattering on the total path length and is nonlinearly dependent on both the absorption and scattering properties of the medium.

Recently, an extension of the YM model to optically thick samples has been developed [11].

A new expression for the scattering coefficient of light s_∞ was proposed in this regime:

$$s_\infty \propto \sqrt{\frac{e^{-\frac{1}{\sqrt{1+\alpha^2}}}}{[4(A_{KM}^2 + 2A_{KM})]^{\frac{1}{4}}}}, \quad (8)$$

and its validity for strongly absorbing paper samples was demonstrated in Ref. [11].

In this way, the cellulose absorption coefficient can be recovered up to the highly absorbing ultraviolet region from nondestructive reflectance measurements.

In order to measure the absolute diffuse reflectance of paper samples, an experimental setup using an integrating sphere to illuminate the sample with diffuse radiation and to collect radiation reflected over all angles is necessary in order to apply the KM and YM models [9, 11].

Recovering the optical absorption of cellulose fibers is an important step in order to make comparisons with theoretical simulations and obtain information on the chemical degradation of paper induced by aging and causing its yellowing.

S

Theoretical Optical Spectroscopy

Because of the complexity of paper structure, it is not possible to associate optical absorption peaks to specific oxidized groups by a simple comparison with spectra of reference compounds. Due to the important role of the local chemical environment around the oxidized groups, which can result in geometrical distortions and in energy level shifts, their optical properties can be strongly modified making the above comparison unreliable.

A possible pathway to tackle this problem relies on the comparison of the experimental optical spectra with those obtained by computational simulations based on ab initio methods. These allow to simulate physical properties of materials (such as the geometry, the electronic and optical properties, and so on) starting from the microscopic knowledge of their chemical structure. The optical spectra of paper can therefore be calculated using these techniques and compared with the experiments.

Density functional theory [12], for which Walter Kohn was awarded with the Nobel Prize for Chemistry in 1998, represents by now the state of the art of the first principle methods for the calculations of the geometry and all ground-state properties of any system of interacting electrons and ions.

Within DFT the many-body problem:

$$H_e \Psi = \left[-\sum \frac{1}{2} \nabla_i^2 - \sum \frac{Z_A}{r_{i,A}} + \sum \frac{1}{r_{i,j}} + \sum \frac{Z_A Z_B}{R_{A,B}} \right] \Psi(r_1, r_2, \dots, r_N) = E \Psi(r_1, r_2, \dots, r_N) \quad (9)$$

is mapped into a system of noninteracting particles that can be solved by finding the solution of a single particle equation (known as Kohn-Sham equation):

$$\left(-\frac{1}{2} \nabla^2 + V_{\text{ext}}(r) \right) \psi_{n,k}^{KS}(r) = \varepsilon_{n,k}^{KS} \psi_{n,k}^{KS}(r) \quad (10)$$

with, by construction, ground-state electron density identical to the one of the interacting system. In Eq. 9, the ions are considered as fixed point charges with electric charges Z_A and Z_B and respective distances $R_{A,B}$, while electrons are instead considered by quantum mechanics governed by the Schrodinger equation. The elements of the Hamiltonian reported in Eq. 9 are in order: the electronic kinetic energy, the electron-ion interaction, the electron-electron interaction, and the ion-ion interaction.

In Eq. 10, V^{ext} is the external potential and includes the electron-ion interaction, the Hartree potential, and the so-called exchange and

correlation potential, which is defined as the functional derivative of the exchange and correlation energy. The external potential is a functional of the electronic charge density and is constructed in such a way that the noninteracting electron system ground-state charge density is the same as the one of the interacting electron system. The total energy of the interacting system is univocally determined by its ground-state electron density, as demonstrated by the Hohenberg and Kohn theorem within the density functional theory [12]; hence the knowledge of its ground-state density (through the solution of Eq. 10) should in principle give us access to the ground-state energy. The problem arises from the fact that the expression of the total energy as a functional of the electron density is unknown. In particular, the analytical form of one important ingredient, the so-called exchange and correlation energy, is unknown, and as a consequence, also the exact exchange and correlation potential is unknown. Approximations have to be used, such as LDA, GGA, and so on. For this reason, even if DFT is a formally exact theory, the total energy of the interacting system can only be evaluated within approximations. The structure and the main theorems on which the DFT is based can be found in several references (see, e.g., Ref. [12]), as well as the several approximation used to evaluate the exchange and correlation potential.

If ground-state properties can be successfully calculated within DFT, in order to compute optical spectra, which are an excited state property, a different approach has to be used. Time-dependent density functional theory (TDDFT [13]) can be considered as an extension of DFT to time-dependent external potentials and a (in principle) rigorous approach for optical spectra calculations.

Methods based on TDDFT allow to calculate the response function that relates the time-dependent external electric field to the response of the electronic charge density inducing a polarization in the material:

$$\chi(r, r'; t - t')_{\alpha, \beta} = \frac{\partial P_\alpha(r, t)}{\partial E_\beta(r', t')} \quad (11)$$

Here, P is the polarization of the material and E the external electric field. The response function χ is related to the optical properties of the material (dielectric function, optical absorption, reflection, and so on). TDDFT allows to obtain the response function of an interacting electron system from the one calculated for a noninteracting electron system (χ_0) by the following recursive Dyson-like equation:

$$\chi = \chi_0 + \chi_0(v_c + f_{xc})\chi \quad (12)$$

where v_c is the Coulomb potential and f_{xc} is the exchange-correlation kernel. The kernel is a functional of the electronic charge density. Its exact analytical expression is not known, and several approximations are used to estimate it (as, e.g., ALDA, APBE, ABLYP, and other usually taken from DFT with adiabatic approximation) [13].

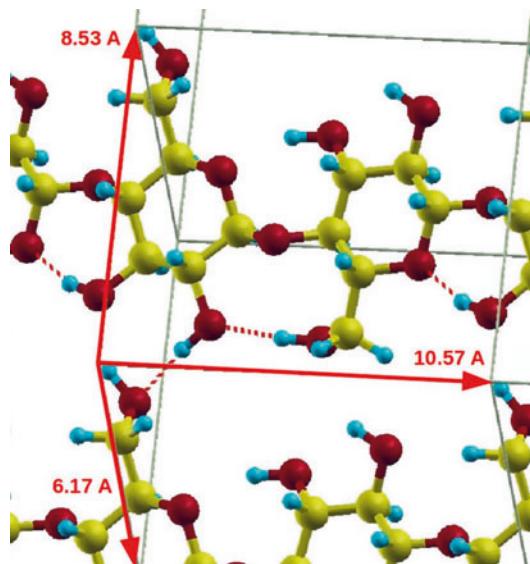
TDDFT calculations of optical spectra in complex systems represent by now the state of the art especially in molecules, nanoclusters, and in biological systems.

Theoretical Optical Spectra for Oxidized Cellulose

In order to interpret the experimental optical absorption spectra and recover chemical information at the nanoscale, the knowledge of the optical absorption spectra of the various chromophores, which are supposed to be responsible for aged paper yellowing, is needed.

Density functional theory and time-dependent density functional theory methods can be applied to the study of cellulose, as reported in Refs. [6, 14]. In these works, as a structural model for cellulose, an infinitely extended crystal of cellulose has been chosen as an approximation of cellulose crystalline domains. Cell parameters of cellulose crystal were obtained by X-ray diffraction data [2–4] and refer to the monoclinic crystallographic phase called cellulose-I_B, which is the most common polymorph for higher-plant cellulose [3, 4] (see Fig. 2).

Cellulose with several forms of oxidized groups has been considered in the simulations.



Spectroscopy of Ancient Documents,

Fig. 2 Monoclinic crystallographic phase of cellulose-I_B, with cell parameters taken from X-ray diffractions [3, 4] and atomic coordinates inside the periodic cell calculated by DFT. Light-gray (yellow), dark-gray (red), and small (blue) spheres represent, respectively, carbon, oxygen, and hydrogen atoms. *Dashed red lines* indicate hydrogen bonds

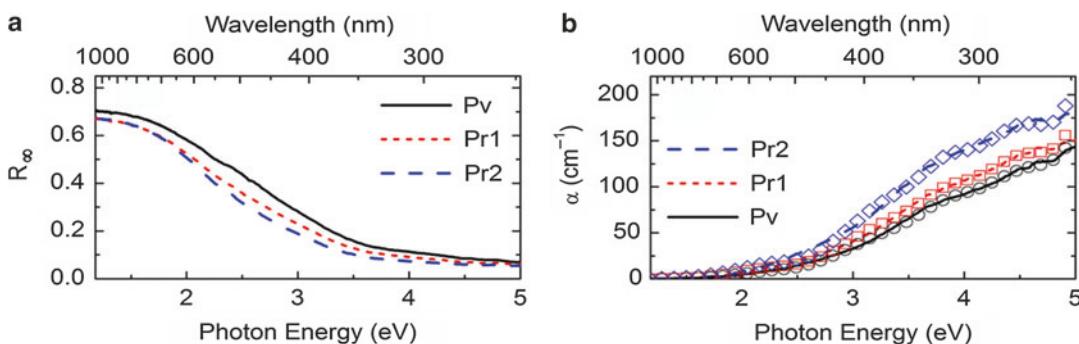
Simulated theoretical optical spectra can be compared to experimental ones by a fitting procedure. This comparison allows the characterization and estimation of the quantity of oxidized groups acting as chromophores inside aged paper starting from nondestructive optical experimental measurements.

This approach has been applied to several ancient paper documents, some of which are of great importance for cultural heritage, and has also been applied to the famous Leonardo Da Vinci's self-portrait [14].

Applications

The uniqueness and fragility of ancient pieces of art on paper require a noninvasive and nondestructive diagnostic method such as optical spectroscopy. For this reason, diffuse reflectance measurements have been used to evaluate the degradation of several paper artifacts.

Spectroscopy of Ancient Documents, Fig. 3 Front and back of the Leonardo Da Vinci's self-portrait. Red short dashed, circle blue long dashed, and circle black-solid circle indicate the spots where reflectance measurements were carried out



Spectroscopy of Ancient Documents, Fig. 4 Optical reflectance (panel **a**) of the Leonardo Da Vinci's self-portrait measured in the three spots shown in Fig. 3 and respective absorption spectra derived by using the

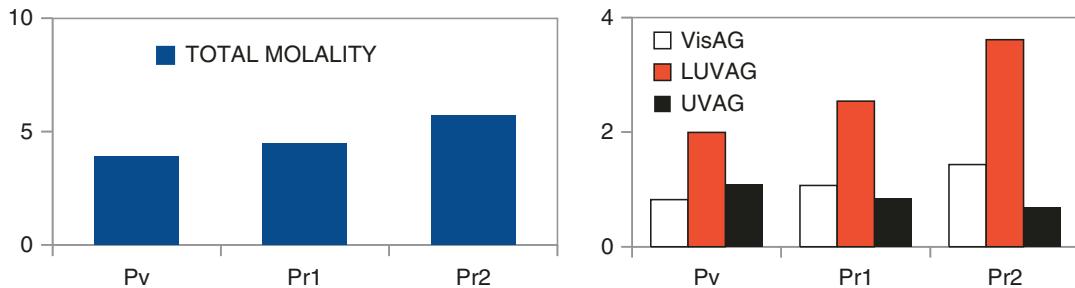
Kubelka-Munk approach described above (panel **b**, lines). Theoretical absorption spectra of spots are also reported in symbols in panel **b**

For example, the diffuse absolute reflectance spectra of the Leonardo Da Vinci's self-portrait were measured in two spots on the front of the sheet (blue and red dashed circles in Fig. 3) and in one spot on its back (black continuous circle in Fig. 3).

The spectra of the artwork were acquired over a wavelength range from 250 to 1050 nm (4.96–1.18 eV of photon energy) (Fig. 4). From the reflectance spectra, the experimental intrinsic absorption coefficient of the constituent cellulose fibers of the paper of the Leonardo self-portrait was recovered by using the extension for strongly absorbing samples (see Eq. 8) of the YM and KM

models. For each oxidized form of cellulose (including ketones, aldehydes, and diketones within the β -D-glucopyranose units), the optical spectra were obtained by TDDFT calculations [6, 14].

To analyze our results, chromophores were divided into three main categories according to their optical response: visible (VisAG), low-ultraviolet (LUVAG), and ultraviolet (UVAG) absorbing groups. The VisAG category contains the diketones. The LUVAG category contains aldehydes and single ketones. The latter are characterized by strong H-bonds with the hydroxyl groups. Although the chromophores of the LUVAG category mainly absorb in the low



Spectroscopy of Ancient Documents, Fig. 5 Total (*left* panel) and partial (*right* panel) concentrations of chromophores measured by spectroscopic approach in the three spots of Fig. 3

UV range, they also display a smooth tail extending into the short end of the Vis region (blue violet). Chromophores classified as VISAG and LUVAG are the ones that are the chief ones responsible for the yellowing of paper. Finally, the UVAG category contains single ketones free of hydrogen bond-type interactions which do not directly give rise to visible damage.

Comparison between experimental and theoretical spectra allowed the calculation of the total concentration of chromophores in the measured spots. Values were found to range between 3.3 and 6.1 mmol/100 g of cellulose, figures similar to those found in ancient samples of a comparable age.

The chromophore concentrations found in the front spots resulted to be about 20–80 % greater than that of the back spot (Fig. 5). This points to the front part of the work having been heavily exposed to external degradation agents. This observation implies that any further exposure of the self-portrait to such agents will increase the difference in optical degradation between the front and back of the sheet.

Comparing the concentration of UVAG, LUVAG, and VisAG chromophores to those found in modern papers which were artificially aged in different environmental conditions [14] highlights the important role played by humidity in the degradation of the Leonardo's self-portrait. In addition to the presence of moisture, degradation products within a sealed housing may have catalyzed further oxidative processes in addition to those which were already underway.

A periodic repetition of the same analysis would provide an ongoing quantitative

assessment of its rate of degradation and would increase our understanding of the inevitable degradation processes that are underway.

Another example of application of optical spectroscopy of paper is the analysis of degradation phenomena in paper used as insulator of higher voltage winding in power transformers [11]. This kind of paper is made of pure cellulose (without lignin and fillers). The environmental conditions within power transformers are very harsh, with temperatures around 150 °C.

Therefore, the level of oxidation is one order of magnitude larger than that found in ancient paper samples aged during several centuries. The total concentration of chromophores is about 18 mmoles/100 g of cellulose.

The possibility to quantify the level of visual degradation of these paper materials by a nondestructive and noninvasive approach and observe its evolution in time represents invaluable information for conservators and restorers. Moreover, paper has several industrial applications; therefore, the field of application of this approach can be extended beyond cultural heritage.

Optical spectroscopy has a wide range of applications for cellulose-based materials, like paper, textiles, and other manufactured products of great industrial and cultural interest. It can also be extended to other strongly absorbing inhomogeneous materials.

References

1. Hunter, D.: Papermaking: The History and Technique of an Ancient Craft. Dover, New York (1978)

2. Krassig, H.A.: Cellulose: Structure, Accessibility, and Reactivity. Gordon and Breach Science, Singapore (1993)
3. Klemm, D., Philipp, B., Heinze, T., Heinze, U., Wagenknecht, W.: Comprehensive Cellulose Chemistry; Volume I: Fundamentals and Analytical Methods. Wiley VCH, Weinheim (1998)
4. Klemm, D., Heublein, B., Fink, H.-P., Bohn, A.: Cellulose: fascinating biopolymer and sustainable raw material. *Angew. Chem. Int. Ed.* **44**, 3358–3393 (2005). doi:10.1002/anie.200460587
5. Zervos, S.: Natural and Accelerated Ageing of Cellulose and Paper: A Literature Review in Cellulose: Structure and Properties, Derivatives and Industrial Uses edited by A. Lejeune, and T. Deprez, p. 155. Nova, New York (2010)
6. Mosca Conte, A., Pulci, O., Knapik, A., Del Sole, R., Lojewska, L., Missori, M.: Role of Cellulose Oxidation in the Yellowing of Ancient Paper. *Phys. Rev. Lett.* **108**, 158301 (2012)
7. Kerker, M.: The Scattering of Light and Other Electromagnetic Radiation. Academic, New York (1969)
8. Kortum, G.: Reflectance Spectroscopy (Principles, Methods, Applications). Springer, Berlin (1969)
9. Hubbe, M.A., Pawlak, J.J., Koukoulas, A.A.: Paper's appearance: a review. *Bioresources* **3**(2), 627–665 (2008)
10. Yang, L., Miklavcic, S.J.: Revised Kubelka–Munk theory. III. A general theory of light propagation in scattering and absorptive media. *J. Opt. Soc. Am. A* **22**, 1866 (2005)
11. Missori, M., Pulci, O., Teodonio, L., Violante, C., Kupchak, I., Bagniuk, J., Lojewska, J., Mosca Conte, A.: Optical response of strongly absorbing inhomogeneous materials: Application to paper degradation. *Phys. Rev. B* **89**, 054201 (2014)
12. Parr, R.G., Yang, W.: Density-Functional Theory of Atoms and Molecules. Oxford University Press, New York (1989). ISBN 0-19-504279-4. ISBN 0-19-509276-7
13. Runge, E., Gross, E.K.U.: Density-functional theory for time-dependent systems. *Phys. Rev. Lett.* **52**(12), 997–1000 (1984). doi:10.1103/PhysRevLett.52.997. Bibcode:1984PhRvL..52..997R
14. Mosca Conte, A., Pulci, O., Misiti, M.C., Lojewska, J., Teodonio, L., Violante, C., Missori, M.: Visual degradation in Leonardo da Vinci's iconic self-portrait: A nanoscale study. *Appl. Phys. Lett.* **104**, 224101 (2014)

Spider Silk

Fritz Vollrath

Department of Zoology, University of Oxford,
Oxford, UK

Synonyms

[Silks of Spiders as model Bio-polymers](#)

Definitions

Silks are animal fibers (or more rarely ribbons or sheets) of proteinaceous biomaterials that are, by definition, extrusion spun [1]. While the evolutionary origins and taxonomic placement of silk feedstocks can differ widely across the arthropods, filaments can be surprisingly similar [2]. *Capture silks* are sticky materials deploying either nanoscale filaments or aqueous glycoprotein glues that have evolved from dry silks [3].

Outline

Silks are fascinating biological products and have evolved several times independently in the arthropods. Spiders and moths are the best-known and best-studied of silk spinners, but there are others ranging from mites to bees [2]. In each taxon the diversity of silks has evolved in only one ancestor but then radiated quickly (over millions of years) into many different types fit for the various purposes required by the animal – be it integration into a cocoon composite or use as a single safety line. Spiders are unique among silk spinners in that an individual has a veritable armory of silk glands with each silk being tailored to a specific use. Accordingly the silks of spiders show a surprising diversity ranging from tough dry threads to soft wet fibers and from sticky aqueous glue to adhesive nanoscale filaments [4]. This very diversity provides a window into the interaction of a silk's molecular structure with the animal's ecology and thus delivers deep insights into the material's function-structure relationship on the macroscopic, organismic scale. In addition, and of relevance to the subject of nanotechnology, such diversity also offers examples for function-structure relationships on the micro- and nano-scale, and these are the dimensional domains explored here.

However, in order to better understand the scope of the paradigm, and to better appreciate the range of possible solutions, it may be helpful to first briefly outline the biology of the system, and in this way examine the constraints as well as demands imposed on the material “silk”. This

more generic introduction will be followed by specific lessons learned so far from the study of spider silks and from comparisons to the commercially much more important mulberry and wild insect silks. The essay concludes with the briefest of outlooks into the future of silk as a model material for academic investigations into nanoscale processes.

Key Principles

Importantly, silks have evolved to function in the dead state, i.e., detached from the animal and typically after having been dehydrated and denatured [1]. This means that silks provide a model in which we can study – under realistic functional conditions and in great detail – the thermomechanical properties of a biological material. Indeed, there are very few (if any) other biological materials that offer comparable access to in-depth investigations under fully natural conditions. This feature of “natural function even in the fully detached state” provides detailed as well as biologically realistic data about a given silk’s material properties. By applying general biological knowledge one can speculate about property–function relationship ([► Plasticity Theory at Small Scales](#)). But, in order to understand a silk’s structure and its potential relationships with the material’s properties, one requires more than biological insight, i.e., structural data are needed. Such data would come from fine-grained structural analysis such as NMR, X-ray, and neutron-scattering as well as Raman, IR, and CD/LD spectroscopy as well as data on gene and molecular sequences and insights into biological constraints such as the overbearing role of water in biological systems and the evolutionary history of the organism and the silks under investigation. Not surprisingly, while a good understanding of a specific silk’s material properties is emerging, the links to function and structure are still rather weak [5].

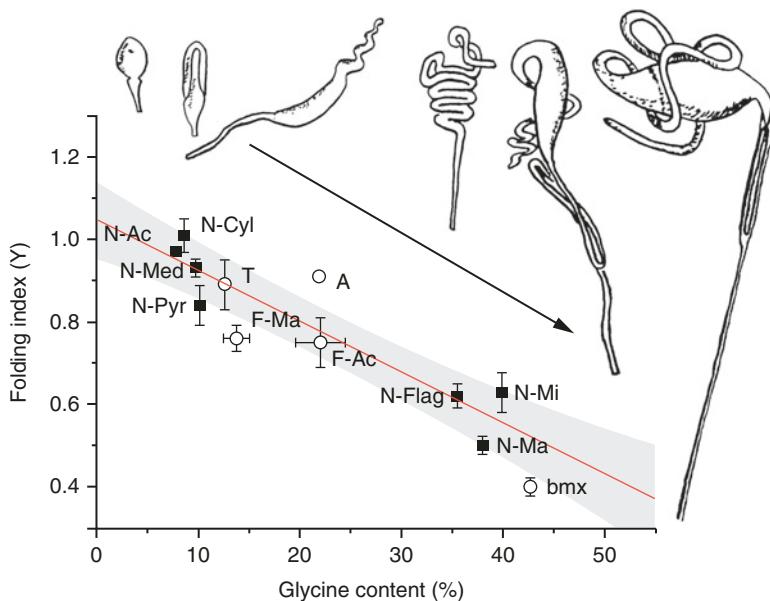
Accordingly it might be more realistic, at this stage, to view the study of silks as research aiming to link *properties* to *function* and to *structure* – rather than taking the more common

approach of exploring a “straight” structure–function relationship. This view notwithstanding, with modern molecular techniques (ranging from sequencing to scattering and modeling) the protein and bond interactions that underlie the functional properties are rapidly being uncovered, and a composite picture is beginning to emerge with clear-cut hypotheses that are ready for rigorous testing [6].

Why Study Spider Silks?

The threads of the commercial mulberry silkworm *Bombyx mori* are the fibers spun by the larva of a lepidopteran moth as part of its cocoon composite constructed to shelter the pupa during its metamorphosis. These fibers have been the mainstay of a key textile industry for over 6000 years and as such have seen thousands of years of ever more sophisticated R&D. However, as the worm spins from converted salivary glands through the mouth, it can cut the thread, which (as we will see) poses certain problems for studying the extrusion process itself. Moreover, lepidopteran silks have evolved for integration into multilayered cocoon complexes and thus have never been optimized for single-fiber strength or toughness but for the specific qualities required of threads in a composite.

Spider silks, on the other hand, have evolved for a much wider range of applications (Fig. 1). The best-studied of these silks, the dragline fibers of the *Major Ampullate Glands*, are optimized for effective and efficient deployment as single-thread safety lines and for integration as key structural members in a web’s open-mesh network. In both applications, these threads’ primary function is always the mitigation of kinetic energy, be it for body support or prey impact and restraint [8]. Many spider silks, therefore, are excellent examples for naturally evolved protein (bio)materials optimized for toughness. Moreover, spiders spin from modified, abdominal leg glands through specialist spinnerets, far away from any cutting mouthpart. Hence spiders can be “silked” easily and under controlled experimental conditions [9].



Spider Silk, Fig. 1 Relationship between the Glycine content of a silk and the “folding index” γ of its proteins in the prespun liquid stage. γ is taken to be indicative of the protein intrinsic disorder. The correlation and model are taken to quantitatively explain the structure-function relationship by describing the molecular conformation i.e., β -sheet propensity. The data suggest that, in order to achieve specialization and performance, silks require higher structural flexibility at the expense of reduced stability and increased conversion energy. A γ value near 1 would denote helix-type folding, while γ values <0.5 would signify mostly unfolded chains having been calculated from the ratio of the circular dichroism spectrum bands at 208 and

220 nm (at 20 °C). The arrow shows the direction of gland evolution and the insets depict schematically the overall gland shape (not to scale). *Nephila edulis* (golden orb spider) major ampullate (N-Ma), minor ampullate (N-Mi), flagelliform (N-Flag), cylindroidform (N-Cyl), aciniform (N-Ac), pyriform (N-Pyr), median (N-Med); *Kukulcania hibernalis* (Filistatidae) major ampullate (F-Ma) and acinous (F-Ac). *Antrodiaetus unicolor* (Antrodiatidae) single type glands (A), and *Aphonopelma chalcodes* (Theraphosidae) acinous (T), as well as the commercial mulberry silk from *Bombyx mori* (Insecta: Bombycidae) (bmx – silkworm silk) (For details see Ref. [7])

Such experimental access allows us to explore the contribution of environmental key parameters (such as temperature, pH, and flow characteristics during extrusion spinning) on the conformation and interactions of supramolecular structures using (even online as the animal spins) high-resolution analytical tools such as X-ray scattering or Raman spectroscopy [9]. Hence much of what is known to date about the molecular structure-function relations in the biological elastomer “silk” originated from studies of spider silks rather than insect silks. Not surprisingly, all relevant studies showed that the key determinant for the silk’s outstanding mechanical properties seems to be its notable structure, which, in turn, relays on the scaling of its semicrystalline morphology [10].

The Importance of Extrusion Spinning

Importantly, the morphology of a silk is the outcome of extrusion processing, not of growth processes. Extrusion “spinning” relies on the refolding of molecules under physical forces such as flow elongation and shear in a tubular duct typically combined with shifts in the chemical environment such as salts and pH changes [11]. Chaperoning molecules might provide additional guidance. Whatever the details of the individual extrusion processing of specific silks, as the fiber is drawn, the solvent water is spontaneously ejected in a self-denaturing solution-gel-solid conversion with the molecules refolding and rearranging themselves from one conformation into another. While the molecular

conformation in the feedstock is adapted to safe storage (sometimes for weeks) in an aqueous solution, the final conformation in the thread is adapted to stability (sometimes for years) against chemical, physical, and biological agents of the environment. The transition between the two conformations typically happens in milliseconds and appears to require very little energy [11]. At present it is known *where* it happens (in the duct), it is known roughly *what* happens (molecular alignment in the flow accompanied by unfolding and hydrogen bond “cross-linking” within and between molecules), but it is not known exactly *how* it happens [12, 13].

One may assume that a silk, be it in the liquid or solid state, adopts shapes and conformations that are dictated by the interactions between polar and nonpolar moieties [13]. A fiber’s macroscale properties are determined by bulk orientations with optimal axial stiffness on the nanoscale. To achieve this, the molecules must be allowed (during the transition from solution to solid) to self-organize into their extended configurations and be accompanied by an efficient intermolecular lock-in. However, in order to maintain flow viscosity, the number of cross-links must be limited. α -helical structures are stable as isolated secondary structures (while β -structures are not) and tend to engage and stabilize interactions with neighboring strands thus forming an intermolecular gelation network. Control of the silk molecules seems to be achieved by a number of reactions. For example, calcium ions stabilize silk proteins at high concentrations but destabilize (and apparently promote β -sheet structures) at low concentrations. Sodium ions also stabilize the silk proteins (important for storage) and are exchanged during the extrusion process by β -sheet promoting potassium ions. Interestingly, similar conditions will produce the formation of amyloids in globular and other fibrous proteins [13].

These (and other) processes during extrusion “spinning” emphasize the importance of a molecular structure and conformation that, while still in the prespun phase, has its future conformations as solid fiber embedded in the amino acid sequence and their accompanying hydrogen bonds

[13]. Importantly for the commercialization of bio-inspired ideas, extrusion spinning is a process that is well established in industrial applications and commercial polymer production, while highly controlled biological growth processes are still outside industrial exploitation (other than via microbial intermediaries). Polymer theory provides a possible tool to explore the various nanoscale structures that self-assemble in the silks during spinning [1]. Ideally, the lessons thus learned can then be used by practitioners to design and manufacture new polymers using novel (and hopefully more sustainable) production methods ([► Bioinspired Synthesis of Nanomaterials](#); [► Self-Assembly](#); [► Sol-Gel Based Nanostructures](#)). After all, silks (and spider silks especially) have many properties that are not only highly eco-friendly but also economically interesting, ranging from full biocompatibility to exceptional mechanical properties.

Mechanical Properties

Spider silks show an enormous diversity of morphologies and mechanical properties, which is not surprising given the great diversity and age of the material [8]. Moreover, spider silks have adapted to a wide range of purposes ranging from “wall papering” burrows to aerial nets, and range in fiber diameters (which is a key parameter of actual strength) from about 10 μm down to a few tens of nanometers [4]. As a rough guide one could take spider silk fibers to have engineering properties (i.e., mechanical properties compensated for fiber dimension) with moduli in the range 1 kPa for a highly hydrated “gel” to about 20 GPa for the stiffest dragline silk, and strengths from almost zero values of yield stress to about 1.6 GPa respectively. In comparison, a decent commercial silk fiber (as used in textiles) would have a modulus and strength of about 10 GPa and 400 MPa respectively [1]. The combination of excellent strength and high strain tolerance gives spider silks their exceptional toughness.

Importantly for the polymer scientist, silks offer a wide range of combinations of stiffness and toughness that might provide, if the

underlying principles are understood, attractive opportunities for the bio-inspired design of novel materials. For example, a dragline of *Nephila clavipes* with a fiber diameter of about 6 μm has a breaking strength of about 1.6 GPa at strains of around 30 %. The estimated yield stress (ca 0.2–0.3 GPa for a 10 GPa modulus) suggests that the spider deploys postyield plastic flow and strain hardening to dampen its movement in a sophisticated combination of viscoelastic properties including the interesting feature of shape-memory [1]. Models and simulations allow the exploration of the molecular processes underlying the stress–strain curves that, for spider silks, can easily be obtained under a wide range of experimental conditions, as outlined earlier.

Modeling Approaches

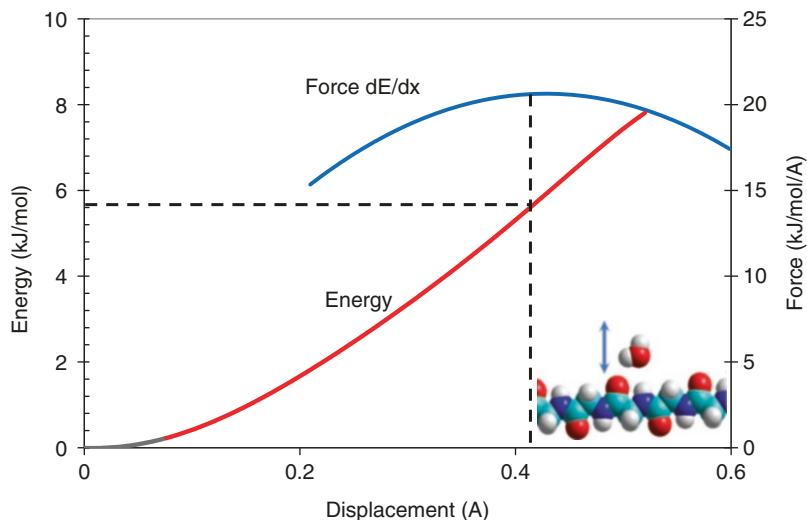
Surprising fits between observation and interpretation can be achieved by suitable models and simulations that set out to predict ab initio the energy interactions between properties and structure in silk feedstocks and fibers (► [Ab Initio DFT Simulations of Nanostructures](#)). For example, quantum mechanics can be used to predict elastic instability energy conditions for hydrogen-bonded interactions between water and the silk proteins. Such energy conditions can then be transformed into parameters such as temperature and stress by using clearly defined relations for zero-point and thermal energies that contribute to the denaturation and glass transition temperatures. A realistic model (e.g., 14) for an energy instability criterion for protein denaturation would have four key stages: (i) calculate the potential energy of water-amide hydrogen bonding as a function of bond length, (ii) calculate the energy at which water becomes free to move away from the amide segments, (iii) calculate the kinetic/entropic energy countering the bond potential energy, and (iv) calculate the probability of water-amide bond dissociation for denaturation as a function of time, temperature, and applied stress.

A specific simulation of stress–strain processes in a spider's dragline silk might serve to

demonstrate here how the general mechanical properties of a dragline silk can be related to internal structure [13]. Molecular dynamics can be used to develop a model that is able to match observed stress–strain profiles with hypothetical “snapshots” of the deforming structures [1]. One insightful model leads to the conclusion that the extraordinary toughness of spider draglines derives from an energy equilibration, as does the postyield strain hardening, which in turn is critical for the strength-toughness balance [14]. This model takes a poly(glycine) “string-of-beads” structure with periodic boundary conditions in an energy-minimized structure to have a density of 1.3 g/cc at 300 K. The purely elastic reference modulus B_e would be given by the cohesive binding energy density with the molecular interactions derived from the Lennard-Jones potential function providing a proportionality constant of 18 for the molecular interactions. Consequently, a cohesive energy of about 40 kJ/mol and a volume of 50 cc/mol for a generic silk peptide segment would give $B_e \approx 14$ GPa. The energy dissipated through the broad secondary relaxation due to hydrocarbon segments would result in an isotropic tensile modulus of about 9 GPa at a Poisson's ratio of about 0.4. The evaluation of such molecular dynamics simulations (see Fig. 2) suggest that the gradual postyield regeneration of hydrogen bonding under strain can absorb large amounts of energy of deformation, which would increase the elastic modulus to its low-strain initial value [16]. There are, of course, also other approaches to molecular modeling, and the field is advancing rapidly.

Of paramount importance for any modeling approach aiming to relate predictions of silk properties to structural features is the reduction of complex protein/polypeptide structures into “ordered” and “disordered” fractions [17]. The ordered domains are more rigid and deform significantly less than the disordered domains, which (like in any amorphous, viscoelastic polymer) would be responsible mainly for energy dissipation, i.e., toughness. Experiments have shown that silk properties are fine-tuned by water acting (as plasticizer) on the disordered fraction where

Spider Silk, Fig. 2 Energy and force as a function of separation distance for a silk water-amide bond, with the elastic instability condition marked by the horizontal dashed line, predicted for the structure shown as an insert using dft simulations (For details see Refs. [13–15])



it binds specifically to polar groups and thus reduces the glass transition temperature. As polar amide-amide bonding determines the glass transition temperature, T_g , (of ca 200 °C in dry silk), increasing hydration of a silk will reduce not only T_g but also modulus and yield stress [10].

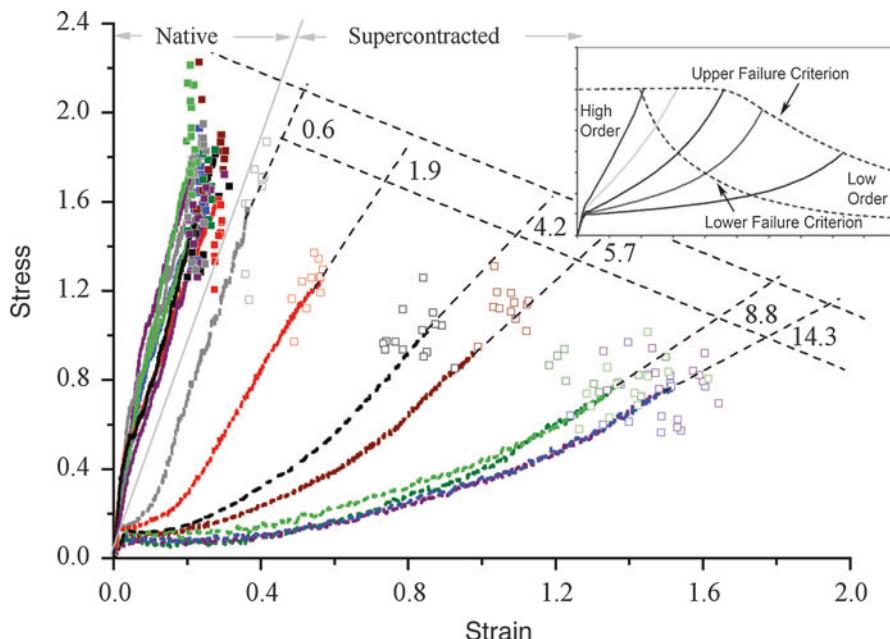
Clearly, as in other organic and inorganic materials, the nanoscale order-disorder organization coupled with degree of hydration are the two key features that define property-structure relationships. An experimental window into these relationships is provided by the comparison of natural silks and silk-like materials prepared from silks that have been dissolved in chaotropic agents to be respun using dehydrating agents. These materials (often called Reconstituted Silk Fibroin or RSF) share molecular sequence but not, apparently, molecular structure/conformation with native silks. Nor do they share mechanical properties, which opens the window for studying the details of property-structure relationships experimentally with the aim of testing specific hypotheses [17].

Functional Properties

Energy is the currency of living organisms. And water is the primary commodity for life and thus the key to understanding biological interactions.

Silk molecules are formed in a fully hydrated state and stored in an aqueous solution. Extraction of water obliges them to go through a sol-gel transition, become immobile, and form the fiber. The rate and ratio of dehydration tunes the mechanical properties of the material via the nature of the hydrogen bonds and their intra- as well as intermolecular interactions.

There is no question that spider silks and their properties can be defined by their sensitivity to water hydration and, indeed, even more powerful chaotropic agents. Some silks swell and shrink when exposed to high humidity or submersion in water while others show no response at all. It seems that the proportion of the amino acid proline is a major contributor to such high water sensitivity – also called supercontraction – reflecting the silk’s observable behavior [18]. Proline, given its shape, would be a side group that brings “disorder” to the folding of the molecular chain. Its presence would thus affect a silk’s material properties by skewing the key ratio of order and disorder (Fig. 3). Interestingly, high-proline content silks not only take up water readily but also become rather “stiff” when redried. Moreover, if such a wetted and redried silk is stretched and restretched repeatedly (i.e., cyclically loaded) it reverts to its original stress-strain behavior. This suggests that there is a reversible change in the order-disorder



Spider Silk, Fig. 3 Representative stress–strain curves of silks with different percentages of proline relating to different nanoscale structural arrangements. The silks were all major ampulate gland MAA *drag-line* silks forcibly reeled from a range of species. The reeling condition had been tuned to produce silk samples with a native breaking strain of 25 % and the silks were tested either in the native (*solid lines/squares*) or the supercontracted state (*dashed lines/open squares*). The proline content and ordered fraction of

supercontracted MAA silks are marked. The *straight black* line defines the upper limit of native MAA silk, where the ordered fraction is 1.0; the *straight gray* line roughly separates the stress–strain curves of native and supercontracted silks (for details see Ref. [1]). The inset shows the calculated mechanical properties of threads with different relationships of order/disorder (For details see Ref. [18])

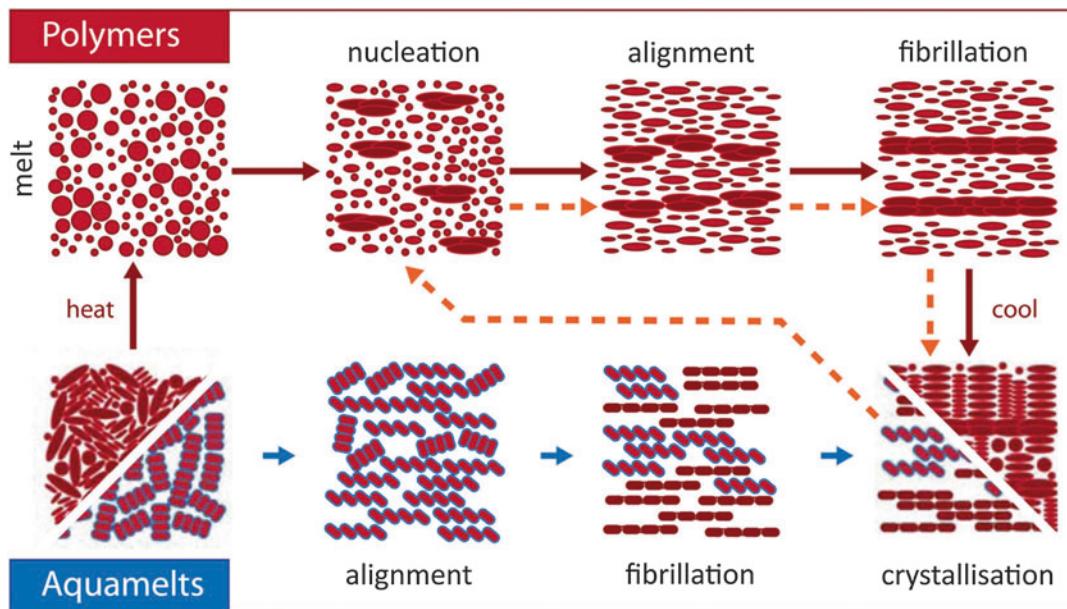
ratio even in the finished fiber, which reflects the level of hydration. It could be argued that the hydrogen bonds of the molecular chain segments around the proline segment are being forced apart under hydration only to be fixed into new positions during dehydration. Rehydration would allow water molecules to reaccess the protein opened up at the proline site and assist in the formation of new hydrogen bonds. This would relax the mechanical strain “frozen” into the material and open the fiber for ingress of water and the exhibition of supercontraction with its accompanying changes in mechanical properties.

Clearly, water in its interaction with the different motifs in the molecular chain of the silk protein can control a silk’s material properties – presumably via the hydrogen bonds at occupied and potential bond sites.

The Importance of Spinning

Generically speaking, the “typical” spider silk consists of two proteins, one rather large (in hundreds of kD) and one significantly smaller. There are notable exceptions, such as the silks of bees, which have three or four smallish proteins (in tens of kD) or the silks of some midges, which seem to consist of hundreds of even smaller proteins [2]. Together, these and some of the unusual spider and lepidopteran silks should provide important test grounds for hypotheses (derived from the “standard” spider silks outlined earlier in this essay) aiming to probe the relationships between structure, function, and properties of silks in general.

The principal hypothesis of nanoscale order–disorder as key to silk properties centers around the importance of a number of processes



Spider Silk, Fig. 4 Fibre formation in polymers (red arrows) and their proposed subclass aquamelts (blue arrows). Horizontal arrows depict shearing steps with length corresponding to the relative amount of work required for fibrillation. Colored objects represent components responsible for fibrillation: long chain molecules (their radius of gyration or shape) in polymers and proteins

(secondary and tertiary structures) in aquamelts with differences in shade corresponding to phase differences. The blue outline of aquamelts represents the outer shell of hydration of the protein. Orange arrows represent the path reconstituted silk follows in order to be reprocessed into a fiber (For details see Ref. [19])

during the formation of the fiber. Indeed, the observable nanoscale structures and interactions appear highly evolved and, by conjecture and experiment, seem to confer on silks the impressive properties that, from the human standpoint, are highly desirable (Fig. 4). Importantly, the molecular interactions have a fixed component (i.e., genetically transferred) as well as a flexible component (i.e., environmentally responsive). Thus, to use a human analogy, a spun silk fiber embodies both nature (the genetic blueprint) and nurture (the milieu during maturation).

Structure (having both a fixed and a flexible component) provides the opportunity to experimentally probe the interactions between structure, properties, and function on the molecular and supramolecular, i.e., nanoscale, level. Firstly, sequencing the silk genes exposes the genetic blueprint and gives the exact positioning of the amino acids that make up the side chains of all silk-polymer macromolecules. Secondly,

extracted liquid silk precursor (feedstock or dope) from the gland of the animal allows examination of molecular conformations in that prespinning state. Thirdly, molecular conformation in the spun fiber state can be studied in the fiber. Importantly, it is also possible to examine the molecular transformations from dope to fiber, although these kinds of study are much more difficult than the others given the small dimensions and quick timescales involved [20, 21].

But it is during extrusion where the protein molecules self-assemble into the complex containing the correct intra- and intermolecular assemblies [11]. Clearly, parameters associated with flow through very fine ducts are important, and not surprisingly given this constraint, the block copolymer molecules of the silk proteins display characteristics of liquid crystals. However, it is still debated whether they really are true liquid crystals or semicrystalline complexes behaving as such [1]. However this may be,

	Soft / Disordered	Hard / Ordered
Spidroin I	<i>GQG GYG GLG SQG A GRG GLG GQG A GAAAAAAAGG A - (G X G)₇ -</i>	α -helix <i>β-sheet</i>
Spidroin II	<i>GPGGY GPGQQ GPGGY GPGQQ GPGGY GPGQQ GPSGPGS AAAAAAAA - (GPGGX)₇ -</i>	<i>random coil</i> <i>β-sheet</i>

Spider Silk, Fig. 5 Schematic of the two principal proteins identified in a benchmark spider dragline silk outlining the link between sequence and properties.

Amino acid codes are *A* alanine, *G* glycine, *P* proline, *Y* tyrosine, *L* leucine, *Q* glutamine

already during the synthesis of the silk protein the molecular complex takes on the distinct morphology of a tightly folded nanoscale rod. Minute changes in pH imposed during the travel down the duct, in combination with mechanical forces, are exerted by the accelerating flow and elongate the fiber as it moves through the hyperbolic section of the extruder [11]. Water is pressed out as the proteins align and interact. Further down in the duct, which is now tubular, additional mechanisms (about which little is known) impose further interactions that give the thread its final shape, inside and out, as well as determining the mechanical properties, which are of course tightly correlated with internal and external structure.

Molecular Sequences

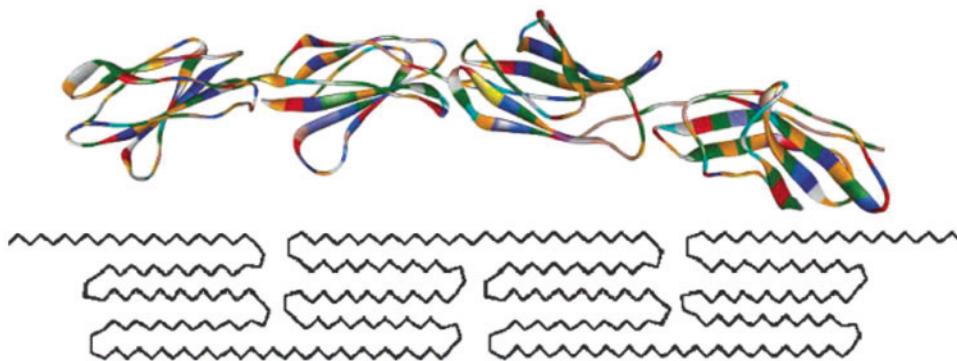
Silks are mostly made up of proteins and contain a range of amino acid combinations. The peptide segments/motifs have a core configuration of -NH-CO-CHR- where R represents 1 of 20 different side groups, which may exhibit highly divergent chemical functionalities (Fig. 5). Spider silk proteins typically have very high fractions of the amino acid glycine, which has the simplest (smallest) of all side groups (G, R = H) and which, because of lack of steric hindrance, allows a wide range of folding conformations. The next important, and still rather simple, peptide groups are alanine (A, R = CH₃) and serine (S, R = CH₂OH) introducing hydrophobicity and polarity, respectively. Finally, there is proline (P, R = CH₂CH₂CH₂), which with its large cyclic and rather rigid side group intrinsically disorders the chains

by twisting them out of their usual highly regular conformer torsional angles.

Importantly, it appears that in spiders the more “advanced” silks seem to have higher ratios of glycine coupled with rather complex spinning glands (Fig. 1). This suggests that molecular flexibility in a silk feedstock requires more control during extrusion; however, the inverse is also possible, i.e., that increased complexity during processing has led to increased molecular flexibility [13].

Proline is another key component of silk nanostructure by forcing the molecular chain that contains it in regular spacings into the string-of-beads morphology mentioned earlier (Fig. 6). This conformation (as yet hypothetical but based on strong evidence) consists of “beads” of beta folds, typically 10 units long and with a chain axis length of about 4 nm each [10]. Such a “string-of-beads” would facilitate processing in the flow field of the narrow extrusion duct [1] by allowing reorientation but not entanglement, thus conferring liquid-crystalline non-Newtonian flow behavior [11].

Clearly, analysis of the primary structure already allows for the formulation of testable hypotheses and related interpretations of the secondary structures. And these are the structures that form most silks’ semicrystalline, nanometer-scale morphology of “hard”, highly ordered domains in a “soft” matrix of more disordered polymer chains (Fig. 3). Degree of hydration is the key as the secondary conformations are held in place by hydrogen bonds connecting the amide groups of the main-chain backbone either directly (hard) or via a water molecule (soft).



Spider Silk, Fig. 6 Silk's nanoscale "string of beads" hairpin folding morphology induced by the proline side chains (For details see Refs. [7, 22])

It must be assumed, although the final verdict is still out, that strong spider dragline fibers have a hybrid structure of about 50/50 hard/ordered to soft/disorderd domains whereby the hard beta sheet folded domains are embedded in a matrix of chains in various degrees of helical conformations thus acting like an amorphous matrix. Indeed, perhaps it is wise to hedge about the minute details of structural function-property relationships until better data has come in from fine-grained analytical spectroscopic techniques such X-ray and Neutron Scattering, Nuclear Magnetic Resonance Fourier Transform InfraRed and Ultra-Violet Spectroscopy, as well as Dynamic Mechanical Thermal Analysis, Differential Scanning Calorimetry, and Thermogravimetric Analysis.

Conclusion

More than 400 million years of evolution will have seen to it that the hierarchical morphology in the patterning and mesophase assembly of spider silk molecular motifs is spatially optimized for each of the many specific functions required for survival of the spider. It has been argued that the evolutionary optimization of energy resources provides a perfect framework for multiscale models probing the intrinsic property-structure relationships of silks. In order to investigate the distribution and exchange of energy at the nanoscale, such a model would deploy the

control of energy storage (strength) and dissipation (toughness) at the molecular level. If correct, then the model should predict the full stress-strain profile of silks to failure, covering the full range right from the strongest dragline threads to the most compliant capture threads in the web. Importantly, it appears that this can be done [10].

The model explaining key aspects of the behavior of finished fibers can be extended to integrate key aspects of the formation of the fiber from the dope [9]. Experimental data demonstrate that "live" (i.e., native) silk dope differs significantly from "dead" (i.e., spun denatured then reconstituted) silk [17]. Studies that theoretically examine this conversion from unspun to spun *native* silk seem to indicate that a silk, once spun, cannot be "unspun" and spun again [20]. This insight, if correct, will have important implications for the goal of producing spinnable silk dopes either by reconstitution from fibers [6] or by protein extraction from microbial expression systems [21]. According to present insights [13], both ways of producing semi-native dope would fatally deconfigure the natural molecular structures; yet these (if the models are correct) would be key requirements for the natural spinning process to function. And that, after all, is one of the most amazing traits of silks and their energetically so efficient and clean production: it uses only water as solvent at ambient pressure and temperature yet produces a biopolymer fiber that can

hold its own proudly in comparison with mechanical properties of man-made polymer fibers.

As outlined in this essay, spiders provide us with a generic class of natural materials, silks, which lend themselves as models for a wide range of other biological elastomers. Unlike these, which naturally perform in the hydrated state, silks have evolved to operate in a wide range of states of hydration, from dry webs and cocoons through to air sacs for underwater spiders, and with the subtle control of tightening sagging webs by supercontraction with dew condensation. Understanding the interaction of biological elastomers with water is a key requirement if we are to produce synthetic biomimetic analogues. Biological functionality, after all, relies on wet engineering coupled with nanoscale dimensions often integrated into structural “hierarchies” of differing degrees of order.

Cross-References

- [Ab Initio DFT Simulations of Nanostructures](#)
- [Bioinspired Synthesis of Nanomaterials](#)
- [Plasticity Theory at Small Scales](#)
- [Self-Assembly](#)
- [Sol-Gel Method](#)

References

1. Vollrath, F., Porter, D.: Silks as ancient models for modern polymers. *Polymer* **50**, 5623–5632 (2009)
2. Sutherland, T.D., Young, J., Weisman, S., Hayashi, C.Y., Merritt, D.: Insect silk: one name, many materials. *Annu. Rev. Entomol.* **55**, 171–188 (2010)
3. Brunetta, L., Craig, C.: Spider Silk: Evolution and 400 Million Years of Spinning, Waiting, Snagging, and Mating, pp. 1–229. Yale University Press, New Haven/London (2010)
4. Vollrath, F.: Spider webs and silks. *Sci. Am.* **266**(3), 46–52 (1992)
5. Fu, C., Shao, Z., Vollrath, F.: Animal silks: their structures, properties and artificial production. *Chem. Commun.* **43**, 6515–6529 (2009)
6. Omenetto, F., Kaplan, D.L.: New opportunities for an ancient material. *Science* **329**, 528–531 (2010)
7. Porter, D., Vollrath, F.: Silk as a biomimetic ideal for structural polymers. *Adv. Mater.* **21**, 487–492 (2009)
8. Harmer, A.M.T., Blackledge, T.A., Madin, J.S., Herberstein, M.E.: High-performance spider webs:

integrating biomechanics, ecology and behaviour. *J. R. Soc. Interface* **8**, 457–471 (2011)

9. Vollrath, F., Porter, D., Dicko, C.: The structure of silk. In: Eichhorn, S.J., Hearle, J.W.S., Jaffe, M., Kikutani, T. (eds.) *Handbook of Textile Fibre Structure*, vol. 2, pp. 146–198. Woodhead Publishing, Oxford/Cambridge, MA/New Delhi (2009)
10. Vollrath, F., Knight, D.P.: Liquid crystal silk spinning in nature. *Nature* **410**, 541–548 (2001)
11. Aldo Leal-Egaña, A., Scheibel, T.: Silk-based materials for biomedical applications. *Biotechnol. Appl. Biochem.* **55**, 155–167 (2010)
12. Dicko, C., Porter, D., Vollrath, F.: Silk: relevance to amyloids. In: Riggaci, S., Bucciantini, M. (eds.) *Functional Amyloid Aggregation*, pp. 51–70. Research SignPost, Trivandrum (2010)
13. Porter, D., Vollrath, F.: The role of kinetics of water and amide bonding in protein stability. *Soft. Matter.* **4**, 328–336 (2008)
14. Holland, C., Vollrath, F.: Biomimetic principles of spider silk for high-performance fibres. Chapter 7. In: Ellison, M.S., Abbott, A.G. (eds.) *Biologically Inspired Textiles*. Woodhead Publishing, Cambridge, MA (2008)
15. Liu, Y., Sponner, A., Porter, D., Vollrath, F.: Proline and processing of spider silks. *Biomacromolecules* **9**, 116–121 (2008)
16. Porter, D.: *Group Interaction Modelling of Polymers*. Marcel Dekker, New York (1995)
17. Holland, C., Terry, E.A., Porter, D., Vollrath, F.: Rheological characterisation of native spider and silkworm dope. *Nat. Mater.* **5**, 870–874 (2006)
18. Spiess, K., Lammel, A., Scheibel, T.: Recombinant spider silk proteins for applications in biomaterials. *Macromol. Biosci.* **10**, 998–1007 (2010)
19. Holland, C., Vollrath, F., Ryan, A.J., Mykhaylyk, O.O.: Silk and synthetic polymers; reconciling 100 degrees of separation. *Adv. Mater.* **24**, 105–109 (2012)
20. Porter, D., Vollrath, F.: Water mediated proton hopping empowers proteins. *Soft. Matter.* **9**, 643–646 (2013)
21. Porter, D., Vollrath, F.: Water mobility, denaturation and the glass transition in proteins. *Biochim. Biophys. Acta (BBA) Proteins Proteomics* **1824**, 785–791 (2012)
22. Vollrath, F., Porter, D.: Spider silk as archetypal protein elastomer. *Soft. Matter.* **2**(5), 377–385 (2006)

Spiders

- [Arthropod Strain Sensors](#)

Spintronic Devices

- [Magnetic Nanostructures and Spintronics](#)

Spontaneous Polarization

► Polarization-Induced Effects in Heterostructures

Spray Technologies Inspired by Bombardier Beetle

Alexander Booth¹, Andy C. McIntosh¹, Novid Beheshti², Richard Walker³, Lars Uno Larsson⁴ and Andrew Copestake⁵

¹Energy and Resources Research Institute, University of Leeds, Leeds, West Yorkshire, UK

²Swedish Biomimetics 3000® Ltd,

iBIC – Birmingham Science Park Aston, Birmingham, UK

³ICON plc, Marlow, Buckinghamshire, UK

⁴Swedish Biomimetics 3000® AB, Stockholm, Sweden

⁵Swedish Biomimetics 3000® Ltd, University of Southampton Science Park, Southampton, UK

Synonyms

Flash evaporation liquid atomization; Liquid atomization through vapor explosion; μMist®

Definition

μMist® is a liquid spray and atomization system inspired by a method that is used as a defense mechanism by several types of bombardier beetle. The liquid atomization, spray formation, and its propulsion/emanation are achieved by causing a liquid in a sealed chamber to flash evaporate on the release of an exhaust valve. This flash evaporation in the confined space of the chamber causes a vapor explosion, the force of which overcomes the surface tension of the liquid, causing it to break down into small droplets and exit the valve with a considerable momentum as a spray.



Spray Technologies Inspired by Bombardier Beetle,
Fig. 1 African bombardier beetle

The μMist® System Development

Initial investigations were inspired by the entomological research of Professor Thomas Eisner into the bombardier beetle's defensive spray system. A typical example of a bombardier beetle from Africa can be seen in Fig. 1. Computer simulations of the spray system, using computational fluid dynamics (CFD), then led to the understanding of the physical principles governing the beetle's unique spray facility, and the μMist® system is based upon these principles. When threatened, the bombardier beetle produces a hot liquid spray from its abdomen, which is ejected at its attacker. Eisner's work focused on how this spray was generated by the beetle. He found that at the rear of the beetle's abdomen was a complex series of glands and chambers in which an exothermic chemical reaction occurred [1]. This chemical reaction heated its constituent liquid to a very high temperature in a reaction chamber, linked directly to an exhaust orifice sealed by a biological pressure-triggered exhaust valve. When the liquid in the reaction chamber reached a certain pressure, this exhaust valve opened, causing the liquid in the reaction chamber to be ejected as a spray.

Spray Technologies Inspired by Bombardier Beetle, Fig. 2 An electron micrograph of the bombardier beetle's spray mechanism (Courtesy of Prof. T. Eisner, Cornell University)



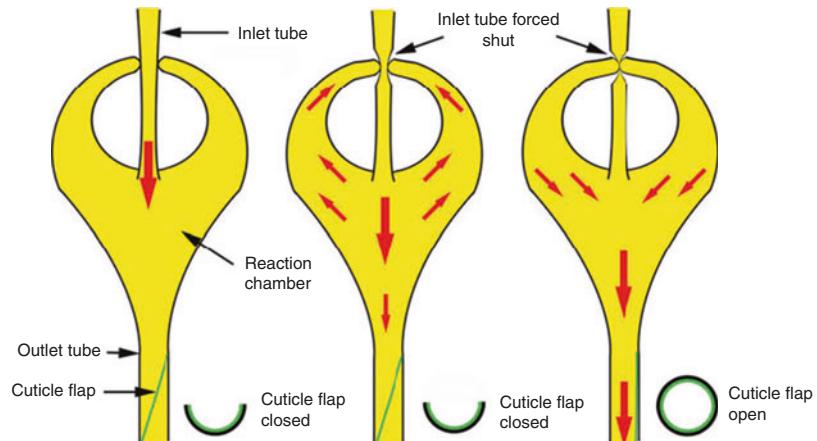
Eisner's work inspired McIntosh and Beheshti [2–5], at Leeds University, to investigate the physical process in the bombardier beetles which generated this spray [2, 3] and found that it is a major advance on established liquid atomization methods. This initial work was funded by the Engineering and Physical Sciences Research Council (EPSRC) and also by Swedish Biomimetics 3000® Ltd. All further work mentioned in this article and all continuing work on the μ Mist® system has been and is currently being supported and advanced by Swedish Biomimetics 3000® Ltd under its V²IO innovation acceleration model. Additional funding has been given by partners for particular projects, including Carbon Connections and the NIHR i4i program.

The initial investigation conducted by McIntosh and Beheshti consisted of a series of Computational Fluid Dynamics (CFD) simulations to better determine the thermodynamic and fluid dynamic processes which were occurring in the chamber of the bombardier beetle. These simulations did not include the associated chemical reaction seen in the bombardier beetle, as in this context it simply functions as a heat source to the chamber. The results of this study showed very good correlation between the simulated behavior of the bombardier beetle's spray mechanism and the behavior observed in the actual beetle [2, 3]. It was these results which inspired the ideas that led to the development of the initial physical μ Mist® system.

The Bombardier Beetle System and Simulation Work

The series of glands and chambers which make up the spray mechanism of the bombardier beetle are shown in Fig. 2, and the most important section is shown in diagram form in Fig. 3. The core of the beetle's spray mechanism is the chamber and the hard cuticle of the outlet tube, which acts as a pressure triggered exhaust valve. Initially the chamber is mostly empty, the cuticle is closed, blocking the outlet tube, and the inlet tube is open. When the beetle decides to spray, reactants are fed through the inlet tube into the chamber in the form of a dilute aqueous solution. These reactants are a toxic mixture of hydroquinone and hydrogen peroxide which react exothermically due to the addition of a catalyst believed to be in the inner surface collagen of the beetle's combustion chamber. This causes an increase of both temperature and pressure in the chamber. As the pressure in the chamber increases, it causes the inlet tube to be pinched shut by the chamber's extremities, forming a sealed volume of liquid (mainly water, plus some quinones from the products of the reaction). The ongoing exothermic reaction further heats the liquid in the chamber above its usual boiling point. Only a small fraction of the liquid is able to evaporate due to both the inlet and outlet tubes now being blocked. Eventually, the chamber pressure becomes high enough that it can force open a cuticle flap which holds the outlet tube shut. This opening occurs very suddenly and is entirely due to the buildup of pressure

**Spray Technologies
Inspired by Bombardier
Beetle, Fig. 3** A diagram of the spray mechanism of the bombardier beetle at three different stages in the ejection cycle



in the chamber. Thus, the cuticle flap is essentially a pressure relief exhaust valve which is passively triggered in the outlet tube. As the liquid in the chamber has been heated to a point high above its normal boiling point, and the opening of the exhaust valve is so sudden, a very rapid evaporation of the liquid occurs, known as flash evaporation. This flash evaporation does not affect the whole volume of liquid in the chamber, so liquid is caused to be forced out of the outlet tube by a rapidly expanding mass of vapor. The evaporation is rapid due to a vapor explosion, as sudden expansion takes place because of the change in bulk volume of water to vapor. This then shatters the remaining water into small droplets and the mixture is ejected with great force, with shear forces also contributing to the break up into a cascade of ever smaller droplets. The size of the droplets produced are inversely proportional to the magnitude of the forces exerted on and in the liquid volume, so larger forces produce smaller droplets as the surface tension and viscosity of the liquid are more easily overcome. As the liquid is ejected from the bombardier beetle's chamber, the chamber pressure drops, which causes chamber physiology of the outlet tube to now be closed again by the collapse of the cuticle valve. Since the chamber is now partially empty the inlet port is open once again and the process repeats itself. This is exactly the same principle used in modern pulse combustors.

This process was modeled by using Computational Fluid Dynamics (CFD) starting the simulation from the moment at which the cuticle valve on

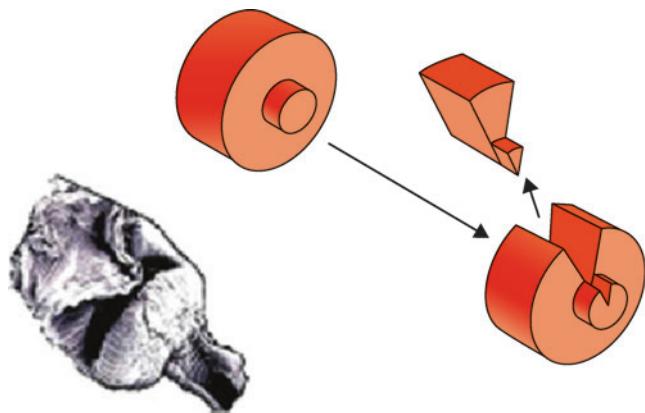
the outlet tube is forced open by the chamber pressure. The chamber is full of liquid (mainly water) already at a suitably high temperature. The inlet tube was not included in the CFD model. The outlet tube however was included in the model and, at the beginning of the simulation, the liquid in the chamber is allowed to flash evaporate. A diagram showing the chamber model used in the simulation can be seen in (Fig. 4) – a slice of the cylindrical chamber is simulated (assuming an axial symmetry within the chamber). The chamber dimensions were such that the volume was similar to that of the bombardier beetle, around 0.1 mm^3 . Good correlation was found between the results of the simulation work and the results of tests on the spray produced by the beetle [3] in terms of characteristics such as spray velocity, throw ratio (i.e. ejection distance divided by chamber length), ejection duration, and mass discharged per pulse of spray. This showed that the method of spray atomization used by the beetle was correctly modeled. With this natural process now understood, the research was taken further by building a scaled-up artificial chamber to develop and thereby assess the suitability of this mechanism as a practical approach to liquid atomization and to conduct further research into the process itself.

Overview of the Physical μ Mist® System

Based upon the simulation work conducted at Leeds University, an experimental demonstration

Spray Technologies

Inspired by Bombardier Beetle, Fig. 4 An electron micrograph of a bombardier beetle's reaction chamber alongside diagram of simulated version used in CFD modeling (Electron micrograph courtesy of Prof. T. Eisner, Cornell University)

**Spray Technologies**

Inspired by Bombardier Beetle, Fig. 5 The pre-prototype μ Mist[®] system in operation



facility (Fig. 5) was built, which implemented the principles of this liquid atomization method as seen in the bombardier beetle. This system is called μ Mist[®]. In a similar way to the bombardier beetle and simulation work, the core of this system is the chamber and valves, as well as a heat source which replaces the heat generated through the bombardier beetle's chemical reaction. Unlike the bombardier beetle and the early simulation work, the exhaust valve on the physical μ Mist[®] system is not opened due to a buildup of pressure in the chamber (that is, passively), but instead, all valves are activated electronically, resulting in an actively controlled pulsed spray. There are also other valves in the system, which control the refill flow into the chamber between ejections, similar to the chamber extremities which control flow into

the beetle's chamber through an inlet tube. Opening or closing these valves and the exhaust valve then controls the flow of liquid as seen in the bombardier beetle and results in atomized spray generation. As with the beetle and CFD simulation work, this is due to the formation of a sealed volume of liquid at high temperature in the chamber just before the exhaust valve is opened. It is important for this high-temperature volume of liquid in the chamber to be sealed to allow the required temperature rise with minimal associated vaporization. This ensures that the chamber liquid temperature is allowed to rise above its boiling point and that a portion of liquid will flash evaporate once the exhaust valve is opened. This produces liquid atomization at much lower pressures than utilized in most liquid atomization devices

(which use pressure atomization – forcing the liquid through small holes at high pressure) and is energetically efficient. The fact that flash evaporation is used to achieve fast ejection of the water and steam, combined with appropriate control systems, means the system lends itself to a fine control of spray characteristics, particularly droplet size. This control is further improved by the decoupling of the exhaust valve from the chamber pressure. In an ideal system, there would be no vaporization of the chamber liquid during the heating phase; however, in practice, some vaporization will always occur, as governed by the equation for vapor pressure (Eq. 1). This is the major cause for the rise in chamber pressure seen in the bombardier beetle, which in that case triggers the opening of the passive pressure relief exhaust valve. Therefore in the bombardier beetle system, this limits the atomization level achieved, since the extent of flash evaporation will be roughly similar at each ejection. The liquid is heated a similar amount and exhausted at a similar time due to the correlation between chamber pressure and chamber temperature. This follows empirical equations [6] such as Eq. 1.

$$\ln P = C_1 + \frac{C_2}{T} + C_3 \ln T + C_4 T^{C_5} \quad (1)$$

Where:

P = Vapor pressure of the liquid

T = Temperature of the liquid

C₁, C₂, C₃, C₄, C₅ = A series of constants depending on the liquid; values for these constants for a range of liquids can be found in [6].

As the exhaust valve on the physical μ Mist® system is not pressure triggered, but actively controlled electronically, the extent of flash evaporation can be controlled to a greater degree than that of the bombardier beetle.

The physical scale of the first pre-prototype of the μ Mist® system is much larger than that of the system found in the bombardier beetle. Whereas the bombardier beetle system has a chamber of approximately 1×10^{-3} m in length, the first μ Mist® pre-prototype system has a much larger

chamber, 0.02 m in length. This level of scaling is not limited to the chamber of the system and extends to other parameters, such as the diameter of the feed and outlet tubes. The process in the μ Mist® system however significantly increases the atomization of the liquid as it exits the chamber. The μ Mist® system can not only be used to atomize very small liquid volumes like those found in the bombardier beetle, but also much larger volumes with a wider range of practical uses. Some key performance parameters are also scaled with the increase in chamber size, such as the throw ratio. A bombardier beetle can throw liquid 0.2 m, using a chamber 1×10^{-3} m in length. This gives a throw ratio of 200 (Eq. 2). Comparatively, it was found that the chamber of 0.02 m in length on the μ Mist® system could also achieve a throw ratio of 200, throwing liquid a distance of up to 4 m [4, 5].

$$R_T = \frac{D_T}{L_C} \quad (2)$$

Where:

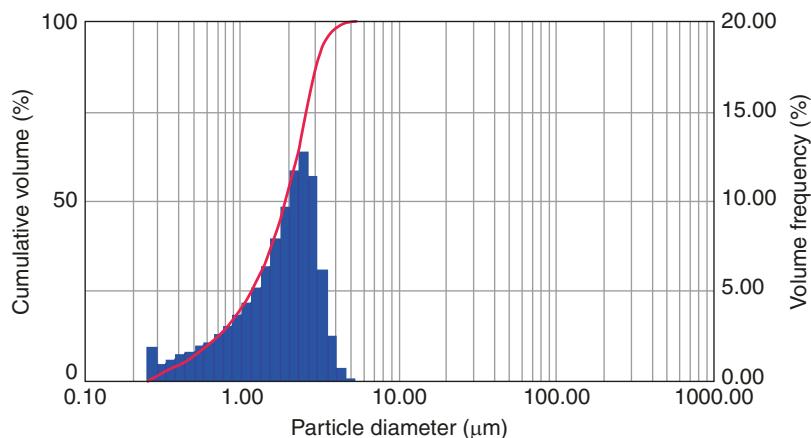
R_T = Throw Ratio (Dimensionless)

D_T = Distance of throw (m)

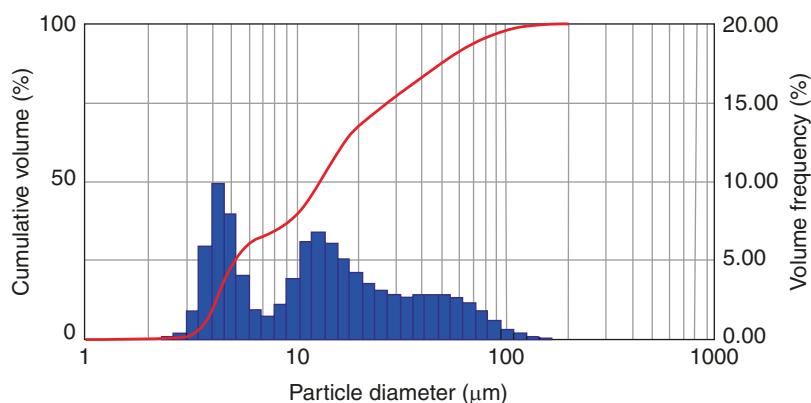
L_C = Characteristic length of Chamber (m)

In summary, despite the differences between the physical μ Mist® system and the liquid atomization seen in the bombardier beetle and the simulation work, the core atomization method is the same. A liquid, heated past its boiling point, and at a constant volume, is suddenly allowed to vaporize through the rapid opening of an exhaust valve. This causes a flash evaporation of a portion of the liquid, which generates a very large force, which then ejects vapor and liquid out through the valve. The flash evaporation is such that as the liquid is rapidly pushed out, large shear forces are generated on and within the volume of the liquid which repeatedly shatter it by overcoming its surface tension and viscosity in a cascade, producing a range of droplet sizes in the emanating spray. Many advantages of the liquid atomization seen in the bombardier beetle, such as high throw ratio, low chamber pressure, and overall energetic

Spray Technologies Inspired by Bombardier Beetle, Fig. 6 Typical droplet size distribution when pre-prototype μ Mist[®] system is producing its smallest droplets



Spray Technologies Inspired by Bombardier Beetle, Fig. 7 Typical droplet size distribution when pre-prototype μ Mist[®] system is producing medium sized droplets



efficiency of the system are also seen in the physical μ Mist[®] system, despite it being on a much larger scale. The μ Mist[®] pre-prototype system also expands upon the liquid atomization method by decoupling the exhaust valve from the chamber pressure, allowing more control over an already versatile system.

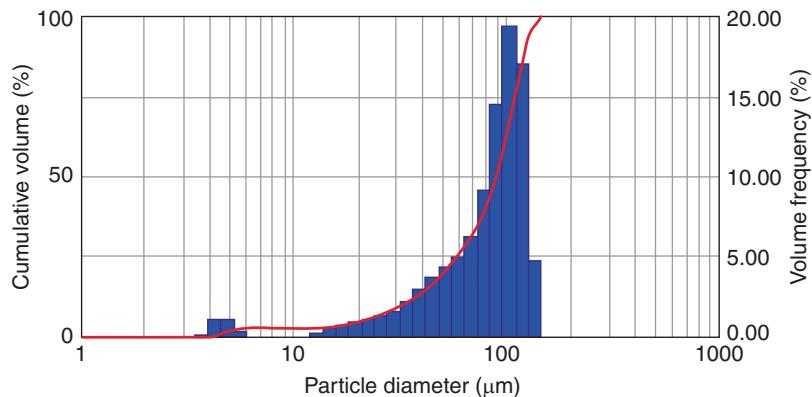
Performance of the μ Mist[®] Spray System

One of the key features of the μ Mist[®] system as a liquid atomization system is that it can be controlled to produce large variation in the characteristics of the spray produced. The type of spray produced is controlled by a number of factors including pressure and temperature in the chamber. Consequently, a range of spray characteristics can be achieved which is much wider than

possible with other liquid atomization systems. Experimental work on the μ Mist[®] system has shown that it can generate a very wide range of spray characteristics such as droplet size distribution, the ejection velocity of the spray, the mass ejection rate, and the temperature of the ejected spray.

One of the most significant characteristic for many spray applications is the droplet size distribution of the spray produced. The volume of droplets in a particular droplet size band is also important. This is usually categorized graphically, such as is shown in Figs. 6, 7, and 8. A usual indicator of size of droplets is given by calculating the droplet size below which the total volume of droplets account for 90 %, 50 %, and 10 % of the overall spray volume. These indicators are termed Dv (90), Dv (50), and Dv (10), respectively. The graphs in Figs. 6, 7, and 8 show typical droplet size distributions recorded during experimental

Spray Technologies Inspired by Bombardier Beetle, Fig. 8 Typical droplet size distribution when pre-prototype μ Mist[®] system is producing its largest droplets



work on the first μ Mist[®] pre-prototype system under different input conditions [4]. These results were recorded using a Malvern Spraytec system, which measures droplet sizes through laser diffraction. The graphs give an idea of the range of droplet sizes that can be achieved by altering basic input parameters of the system. Figure 7 is a good example of bimodality in the droplet size distribution, where the majority of the liquid volume is contained within two peaks, rather than one. This result is significant as some spray applications require two or more ranges of droplet sizes to make up the majority of the spray volume, each of which are designed to perform different tasks and when combined produce a more effective delivery than a unimodal distribution. A μ Mist[®] system would be particularly suited to these sorts of applications. The ability to generate such a wide range of droplet size distribution shown in Figs. 6, 7, and 8 shows that the μ Mist[®] system could potentially provide a single technology which can be used to satisfy a wide range of spray applications.

Applications of the μ Mist[®] Spray System

There are many potential applications for the μ Mist[®] system in the global market of sprays. This is primarily due to the wide range of spray characteristics which can be achieved using this technology and also due to it being much more environmentally friendly than current spray technologies. There are also related energy

efficiencies inherent in the μ Mist[®] system, while using water can replace environmentally unfriendly substances currently in use for aerosols and some other spray applications. Some of the potential applications are currently being actively pursued in industrial development programs led by Swedish Biomimetics 3000[®] Ltd using its V² IO innovation acceleration business model. In this section, a few examples of potential applications for the type of droplet size distributions seen in Figs. 6, 7, and 8 will be examined, showing how the μ Mist[®] system is suited meet to a wide range of requirements.

Fuel Injectors

A very important application of the μ Mist[®] technology is to fuel injectors, and this is one of the potential applications currently being pursued by Swedish Biomimetics 3000[®] Ltd, with initial support from Carbon Connections. The droplet size distribution in Fig. 6 at the lower end of the range of the μ Mist[®] system shows why μ Mist[®] is a potential fuel injector technology. Here it can be seen that very small droplet sizes can be produced, resulting in droplets with a high surface area to volume ratio. This droplet size distribution is particularly suited to combustion applications, such as fuel injection, etc., as the larger overall surface area means that the fuel burns much more efficiently. Current fuel injectors primarily work on the principle of pressure atomization, where liquid is atomized by using high pressure to force it through a small opening to create a fine spray. It is this that generates the stresses on the surface of

the liquid which cause it to break down into small droplets [7], rather than the flash evaporation seen in the μ Mist® system. Generating these high pressures requires a significant amount of energy due to the large pressures used (over 1000 Bar) to ensure proper atomization of the liquid. The energy requirements are large even to produce relatively large droplets. The μ Mist® system would require substantially less energy to properly atomize the liquid, as all that is required is to heat the system to a relatively modest temperature. Additionally, typical fuel injectors produce droplet sizes in the region of 30–80 μm . It can be seen from Fig. 6 that a μ Mist® system is capable of producing droplet sizes far smaller than this, in the region of 1–3 μm with some fluids. This would be a significant improvement over current fuel injection technologies, as not only do the smaller droplets burn much more efficiently, increasing engine efficiency and decreasing fuel consumption, but the more complete burning of the fuel in the engine would also lead to fewer harmful emissions being released into the atmosphere. This is in addition to the much lower energy consumed to atomize the fuel compared to current technologies.

Drug Delivery Systems

Another potential application of the droplet size distribution shown in Fig. 6 – again being pursued by Swedish Biomimetics 3000® Ltd, with support from the NIHR i4i program – is in the design of next-generation drug delivery systems. For many illnesses, the preferred method of drug delivery is directly to the lungs. This is best achieved using small droplet sizes (typically less than 10 μm) that allow the liquid to travel deep into the lungs, where it can be most rapidly absorbed into the blood stream. The inherent efficacy of inhaled therapeutic drugs makes this a key development area. μ Mist® could potentially provide a generic drug delivery system, with one unit being capable of delivering a wide range of drugs, or a personalized medicine device, which is built to satisfy a particular patient's unmet medical needs. Current investigations are also being made to address the pharmaceutical industry's need for innovative drug delivery systems for the administration of

novel compounds, including, but not limited to, peptides and oligonucleotides. This would be a significant step forward in drug delivery technology. Other potential drug delivery applications of the μ Mist® system include needleless injection and nasal drug delivery.

Consumer Aerosols

The μ Mist® system could also have significant impact as a consumer aerosol generator, primarily due to its technically advanced performance, but most importantly due to environmental benefits. Standard spray/aerosol cans, as well as most medical inhalers, generally use Volatile Organic Compounds (VOCs) such as propane and butane to generate the high pressures required to atomize the delivered liquid as it exits the can through the nozzle. It is thought that these VOCs cause environmentally damaging atmospheric compounds such as carbon dioxide and other atmosphere borne reactive species. They are also highly flammable and therefore are a considerable safety risk at the point of use. The μ Mist® system has the capability of delivering consumer spray aerosols as well or even better than the spray systems currently used without the need for possibly harmful dangerous VOCs. The only by-product of this system, besides the active ingredient itself, is water.

Fire Extinguishers and Fire Suppressants

The droplet size distribution seen in Fig. 8 primarily features large droplets (around 100 μm) with a wide overall droplet size distribution (down to around 4 or 5 μm). This sort of droplet size distribution would be well suited to a fire-fighting system as the different droplet sizes achieved both have a significant role in fighting fires. Large droplets, maybe \sim 100 μm , are effective at cooling the fire to a level below its reaction temperature, whereas small droplets evaporate very quickly and move oxygen away from the source of the fire – fire suppression. This range of droplets being produced by one system would allow fires to be extinguished and suppressed efficiently as the fire progresses. Another significant feature of the μ Mist® system when producing droplets in this range is that it has a very high throw ratio,

even for small droplets, and thus targeting is possible. As mentioned previously, the current μ Mist® system has achieved a throw distance of up to 4 m under these conditions, which is a throw ratio with respect to the characteristic chamber length of 200 [4, 5]. With larger chamber volumes, it is envisaged that an even greater throw distance can be achieved, which would be particularly relevant to fire-fighting applications.

In summary, the wide range of spray characteristics and the significantly reduced environmental impact of the μ Mist® spray system allow it to be well suited to many applications, in particular fuel injection, drug delivery, consumer aerosol systems, fire-fighting systems, and fire-suppressant systems. In general, the μ Mist® spray system offers improvements in overall performance, efficiency, and environmental impact. Additionally, the system is economic, being very efficient in energy use – especially in comparison to pressure atomization systems.

Conclusions

The μ Mist® system is a versatile new liquid atomization technology inspired by the unique self-defense mechanism found in the bombardier beetle. The liquid atomization occurs when a sufficiently heated liquid in a sealed chamber is caused to flash evaporate due to the sudden release of an exhaust valve. The flash evaporation mechanism is a new method of providing the large forces necessary to rapidly eject the liquid from the chamber, with shear forces shattering it into a cascade of small droplets. Initial work on the understanding of the bombardier beetle's defense system as a general liquid atomization system was undertaken by McIntosh and Beheshti [2–5] at Leeds University, taking the form of a series of CFD simulations to better understand the key thermodynamic and fluid dynamic principles of the atomization process. This understanding of the thermodynamic and fluid dynamic principles of the atomization process was then used to inspire a design and then construct the first pre-prototype μ Mist® system as a base for experimental work and further research.

The experimental work on the μ Mist® system showed it to be highly versatile, producing a wide range of variation in the resulting spray characteristics. For example, it was shown experimentally that the system could be designed to produce droplets with diameters of anywhere between ~1 and 100 μ m. Other spray characteristics which showed similar levels of variation due to the design of the μ Mist® system included the ejection velocity of the spray, modality cone angle or divergence, the mass ejection rate, and the temperature of the ejected spray.

The large variation in the spray characteristics seen in the experimental work on the μ Mist® system shows it to be adaptable to a wide range of applications, including fuel injection, drug delivery, consumer aerosols, fire fighting, and fire suppression. In some cases, the performance benefits of using the μ Mist® system would create significant efficiency savings and related environmental and economic benefits. In addition, further environmental benefits are possible as the μ Mist® system removes the need for harmful VOCs in those spray systems that currently use them, outputting only water, as well as requiring much less energy than standard pressure atomization systems.

Swedish Biomimetics has a worldwide exclusive licensing agreement with the University of Leeds to research, develop, and to commercialize the μ Mist® platform technology and its various potential applications. Patent pending publication Nos. US11/528,297 and WO2007/0342307 are in place for the μ Mist® platform technology and its application into fuel injection, respectively.

S

Cross-References

- [Bioinspired CMOS Cochlea](#)
- [Biomimetic Flow Sensors](#)
- [Biomimetic Mosquito-like Microneedles](#)
- [Biomimetic Muscles and Actuators](#)
- [Biomimetics](#)
- [Biomimetics of Marine Adhesives](#)
- [Toward Bioreplicated Texturing of Solar-Cell Surfaces](#)

References

1. Eisner, T.: *For Love of Insects*. Harvard University Press, Cambridge (2005)
2. Beheshti, N., McIntosh, A.C.: A biomimetic study of the explosive discharge of the bombardier beetle. *J. Des. Nat.* **1**(1), 61–69 (2007)
3. Beheshti, N., McIntosh, A.C.: The bombardier beetle and its use of a pressure relief valve system to deliver a periodic pulsed spray. *Bioinspir. Biomim.* **2**, 57–64 (2007)
4. Beheshti, N., McIntosh, A.C.: A novel spray system inspired by the bombardier beetle. In: Brebbia, C.A. (ed.), Presented Weds 25th June 2008 as Invited Presentation, Design and Nature IV, Proceedings of 4th International Conference on Design and Nature, Tivoli Almansor, Algarve, 24–26 June 2008. Design and Nature IV. WIT Transactions on Ecology and the Environment, vol. 114, pp. 13–21. WIT Press, Southampton/Boston (2008)
5. McIntosh, A.C., Beheshti, N.: Insect inspiration. *Phys. World Inst. Phys.* **21**(4), 29–31 (2008)
6. Perry, R.H., Green, D.W.: *Perry's Chemical Engineers' Handbook*. McGraw-Hill, New York (2008)
7. Lefebvre, A.: *Atomization and Sprays*. Taylor & Francis, London (1989)

SPS of Carbon Allotropes

- ▶ [Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering](#)

Sputtering

- ▶ [Physical Vapor Deposition](#)

Stable Clusters

- ▶ [Prenucleation Clusters](#)

Stable Pre-critical Clusters

- ▶ [Prenucleation Clusters](#)

Stereolithography

Hyundai Hwang and Yoon-Kyoung Cho
School of Nano-Bioscience and Chemical
Engineering, Ulsan National Institute of Science
and Technology (UNIST), Ulsan, Ulju-gun,
Republic of Korea

Synonyms

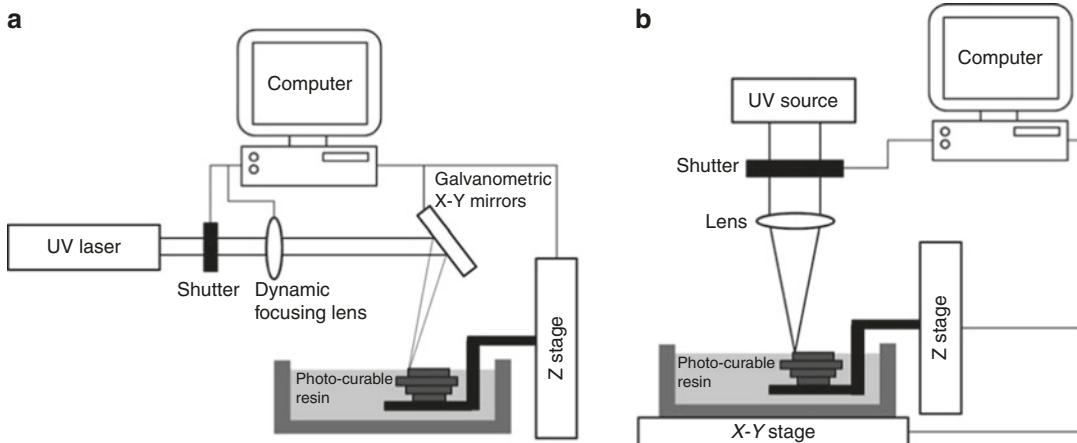
[Microstereolithography](#)

Definition

Stereolithography (SL) is an additive manufacturing process based on photopolymerization of a photo-curable polymer. The solid polymer part can be built via localized polymerization by its selective exposure to light in layer-by-layer format. SL is a kind of solid freeform fabrication (SFF) process for manufacturing solid objects by the sequential delivery of energy or material to specified points in space. SFF is sometimes referred to as rapid prototyping, rapid manufacturing, additive manufacturing, additive fabrication, or layered manufacturing. Microstereolithography or nanostereolithography simply refers to SL which possesses feature resolutions significantly higher than several tens of micrometers or nanometers precision, respectively.

Overview

Stereolithography (SL), which was first introduced in 1981 [1], is a three-dimensional (3D) manufacturing technology based on a layer-by-layer photopolymerization of a photo-curable polymer, so-called resin. A light irradiates or traces in a specific cross-sectional pattern on each layer of the photo-curable polymer, which is initially a liquid state, resulting in the solidification of the resin in the exposed area by the photopolymerization process. After every exposure step for a single layer, the other sliced



Stereolithography, Fig. 1 Schematic illustrations of (a) a classical stereolithography system using a dynamic focusing lens and (b) a microstereolithography system using a *X-Y* stage

two-dimensional (2D) pattern is exposed and polymerized on the below layer by moving the vertical position of the resin chamber or the light. This layer-by-layer 2D fabrication process is repetitively continued until the formation of 3D structure is completed. The performance of SL depends on various factors including the nature of resin, apparatus for fabrication process, strength of light energy, drawing speed of light beam, focusing angle of light beam, depth of focus, and constrained method of liquid-state resin.

Due to the ardent wish of many researchers for reducing the minimum feature of structures that one can manufacture as small as possible, several kinds of stereolithographic techniques, which use much more tightly focused light to initiate polymerization and cure the part in micro- or nano-scale, have arisen. Microstereolithography (MSL), which allows much higher spatial resolution than conventional SL, has been one of the most promising technologies for manufacturing 3D structures in microscale due to its capability for direct fabrication of complicated 3D structures with high accuracy up to 1 or 2 μm . Most of the MSL techniques have been based on the conventional 2D photolithography using ultraviolet (UV) light source and UV curable polymer. Therein, UV exposure for hardening the liquid-state resin is repetitively conducted on each layer

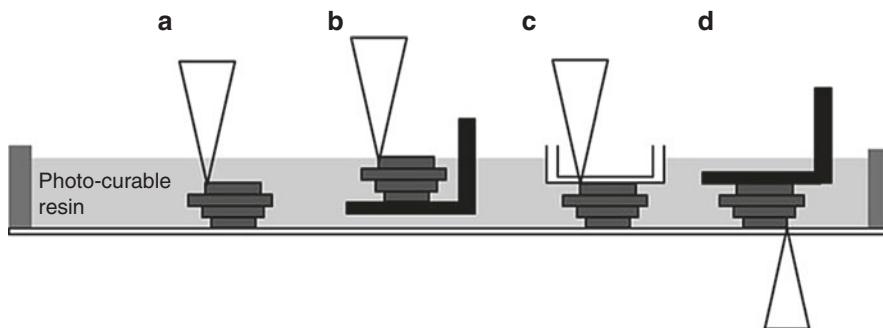
for finally forming 3D microstructures by stacking one layer at a time. Depending on the methods of light exposure, there are two big categories in SL for micro-/nanofabrication: (1) beam scanning-based [2–4]; and (2) image projection-based approaches [5–7]. The former uses a focused light beam tracing a specific 2D pattern over or within the liquid-state resin for each layer, while, in the latter, an image pattern generated by a dynamic pattern generator irradiates only one time for each layer.

Methodology

UV Beam Scanning Stereolithography

In the scanning-based SL process, a very tightly focused laser beam is scanned over or within the liquid-state polymer to produce a 3D solid object. There are several ways to construct a 3D solid part using this scanning-based SL scheme in the perspectives of (1) scanning component – a mirror or an *X-Y* stage (see Fig. 1); (2) light beam direction – from upper or bottom; (3) surface conditions of resin – free or constraint (see Fig. 2); and (4) number of light beam – single or multiple.

With the mirror scanning method, it is very important to maintain the focal point of the light beam over the planar surface of the liquid resin. For this purpose, a dynamic focusing lens



Stereolithography, Fig. 2 Schematic illustrations of four basic configurations for a stereolithography process. Free surface configurations with a light beam projected from upper direction (a) within and (b) over the surface

of the liquid photo-curable resin. Constraint surface configuration with a light beam projected from (c) *upper* and (d) *bottom* directions

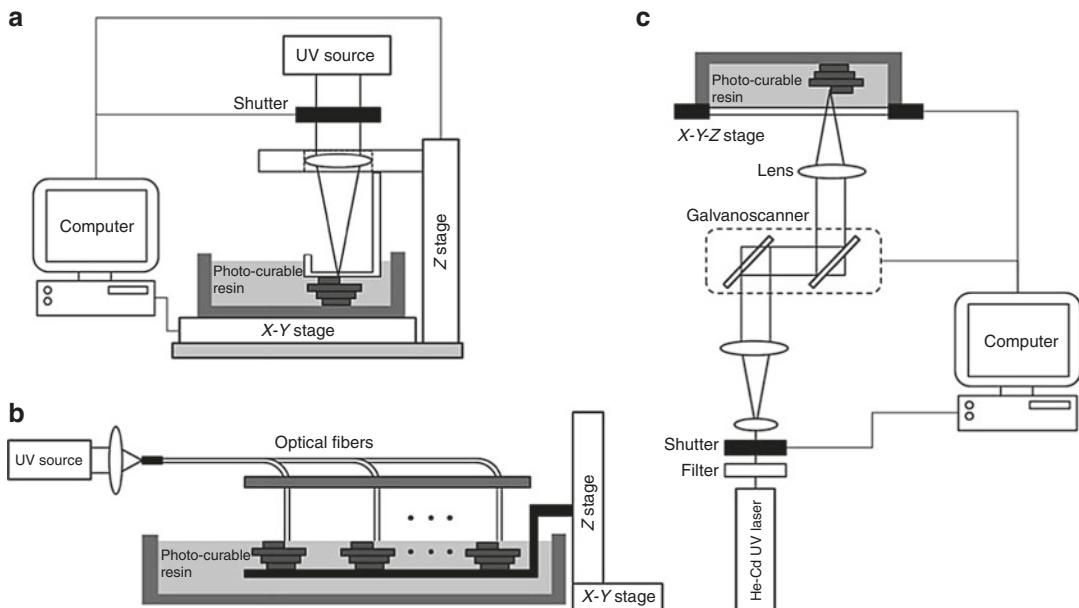
controlled by a computer should be required in a classical SL system (see Fig. 1a). However, it is very difficult to achieve high resolution using this dynamic light focusing. To deal with this problem, not only a Z-stage, but also an X-Y stage for scanning the resin chamber has been applied instead of the dynamic focusing lens (see Fig. 1b). While the chamber containing the liquid resin is translated in the X-Y plane, the focal point of the light beam keeps its position in the plane to achieve the tightest focus as possible. However, the maximum scanning speed of this X-Y stage scanning method is lower than that of the method using a rotating mirror.

Four types of SL system which can be distinguished on the basis of two criteria – the direction of light beam and the surface conditions of resin – are shown in Fig. 2. In the free surface configurations shown in Fig. 2a, b, the surface of the liquid-state resin is free, while the liquid surface is covered by the transparent plate, which allows UV beam passage with little energy loss, in the constraint surface configurations shown in Fig. 2c, d. The free surface configurations are relatively easy to construct, but it is difficult to control the thickness of each layer due to the viscosity of the liquid resin causing a relatively long time to be stabilized. On the other hand, the constraint surface configurations allow more accurate control of the layer thickness by controlling the height of the plate. However, the photopolymerized resin can be often adhered to

the plate and broken in this constraint surface method.

In 1993, Ikuta and his colleagues have first demonstrated the MSL system using this constraint surface configuration as shown in Fig. 3a [2]. They scanned the resin chamber in the X-Y plane and the focused light beam in the Z-axis. They could achieve the minimum feature size of 5 μm in the X-Y plane and 3 μm in the Z-axis. However, it took about 30 min to produce a microstructure, the dimension of which is only 10 $\mu\text{m} \times 10 \mu\text{m} \times 1,000 \mu\text{m}$. To obtain a large number of 3D microstructures at a time using MSL, Ikuta and coworkers reported an advanced MSL system using multiple optical fibers in 1996 [3]. In this system, the light source was tightly focused onto a fiber-optic bundle, and each fiber end was placed in an ordered array for generating a series of exposure point sources in the liquid resin chamber, which is scanned in X-Y-Z directions (see Fig. 3b).

Despite these advances in MSL, there have still been some challenging issues caused by the basic scheme of original SL process, which is a layer-by-layer process. In the layer-by-layer process, the vertical resolution of the polymerized objects is limited by the thickness of the layer piled up. In addition, surface tension of viscous liquid resin can deform and destruct the solidified parts in microscale during the fabrication process. To overcome these limitations, Ikuta and coworkers have developed a more advanced MSL system,



Stereolithography, Fig. 3 Schematic illustrations of several types of microstereolithography (*MSL*) systems. (a) The first-generation *MSL* system based on constraint surface condition, a light from upper direction, and an *X-Y*

stage. (b) Fiber-optic *MSL* system for massive manufacturing in large area. (c) High-resolution *MSL* system using a tightly focused UV laser source irradiating inside the photo-curable resin

in which a UV laser source is tightly focused into the liquid resin, not over its surface, to induce the photopolymerization only at the focal spot of the beam, in 1998 (see Fig. 3c) [4]. Since the photopolymerization occurs only around the focal spot of the beam inside the liquid resin, not the surface, the problems related to the layer-by-layer process, which are mentioned above, can be reduced. They could produce freely moved 3D microstructures with 500 nm resolution using this technique. However, extensive polymerization in the defocused regions, where the beam passes through, is an unavoidable issue associated with this technique.

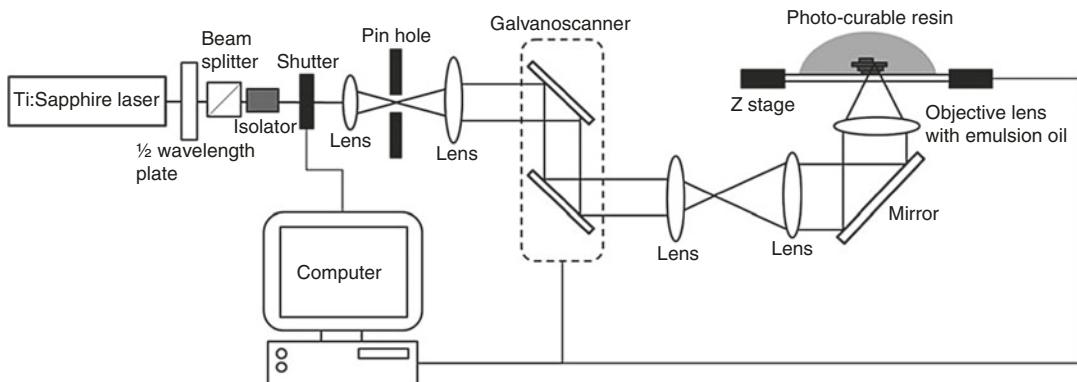
Two-Photon Stereolithography

Two-photon stereolithography (TPS), which is based on two-photon induced photopolymerization, is one of typical SL techniques for the 3D micro-/nanofabrication [8–13]. The TPS technology is based on the polymerization of liquid-state polymer via two-photon absorption (TPA) within very local regions inside the focused high-intensity laser beam tracing in a specific

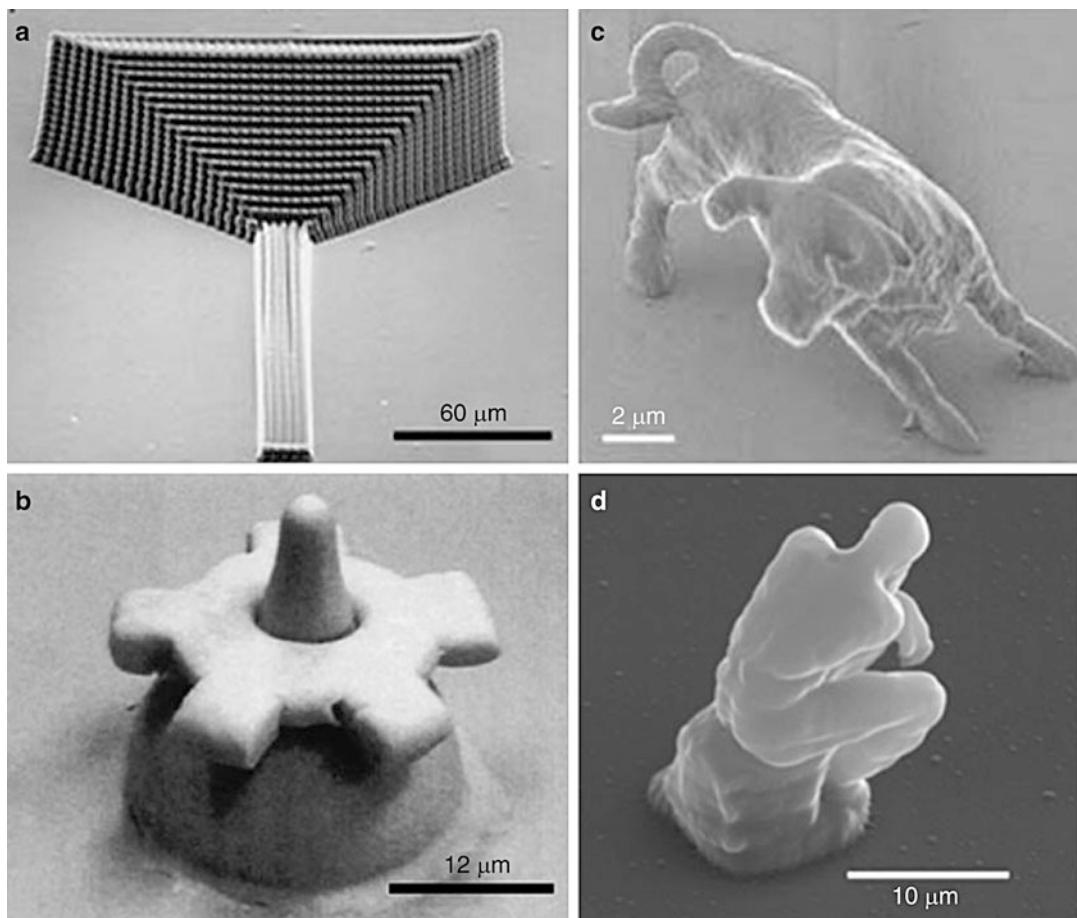
cross-sectional pattern (see Fig. 4). TPA is the simultaneous absorption of two photons by a molecule resulting in its excitation from a low-energy electronic state to a higher energy electronic state. The TPA-initiated polymerization of liquid-state polymer occurs only at a very narrow focal spot of the beam where the light intensity is highest. Thus, the spatial resolution achievable by the TPS is almost 100 nm, which is smaller than the diffraction limit of the beam, while that of other SL systems based on single-photon polymerization is limited by the optical diffraction limit. Based on the fascinated advantages of TPS technology, SL could meet 3D nanofabrication since the 2000s. Not only photonic devices [9] and micromachines [10], but also microsculptures of bull [11] and The Thinker [12] with nanoscale precision have been fabricated using this technology as shown in Fig. 5. To date, the TPS systems have been able to produce the smallest objects based on *MSL*.

Image Projection Stereolithography

Another category of the approaches for *MSL* is the image projection-based one. In this method,



Stereolithography, Fig. 4 Schematic illustration of two-photon stereolithography system



Stereolithography, Fig. 5 Microstructures fabricated by two-photon stereolithography. Various structures including (a) a photonic device (Reprinted by permission from Macmillan Publishers Ltd: Nature [9], copyright (1999)) (b) a micromachine (Reprinted with permission from Ref. [10]. Copyright 2000, The Optical Society),

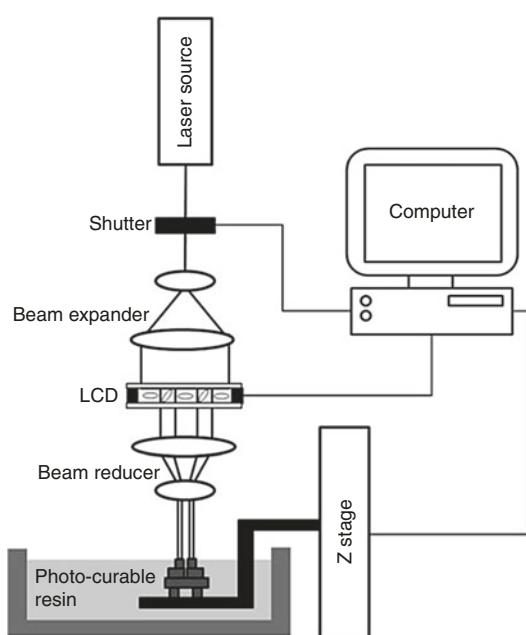
and microsculptures of (c) bull (Reprinted by permission from Macmillan Publishers Ltd: Nature [11], copyright (2001)) and (d) The Thinker (Reprinted with permission from Ref. [12]. Copyright 2007, American Institute of Physics) with nanoscale precision have been fabricated

the photopolymerization and solidification of an object is carried out by irradiating the photo-curable resin with an image pattern in one time for one layer. The image pattern for each layer can be generated by a conventional photo-mask or a spatial light modulator (SLM), which is capable of generating and modulating dynamic image patterns in a continuous way. Typical types of the SLM include a liquid crystal display (LCD) and a digital micromirror device (DMD). Both display devices have a large number of pixels, which are individually controllable. The image pattern generated through the LCD or DMD, which are transmissive and reflective types, respectively, is focused by a lens onto the liquid resin surface. By continuously changing the image pattern using a computer, the resin chamber is scanned in the Z-axis to construct a 3D part (see Fig. 6). Bertsch and colleagues have first demonstrated the image projection MSL system using an LCD as the dynamic pattern generator in 1997 [5]. Since the mid-2000s, several systems using a DMD have been developed and commercialized because the DMD provides better performance – fill factor of

each pixel and reflectivity – compared to the LCD [6, 7]. These image projection-based approaches take much shorter time to fabricate each layer because a large area of liquid resin can be simultaneously exposed in a pre-designed 2D pattern differently from the beam scanning methods, which scan all the area with a tightly focused light. However, these techniques still have the problems due to the layer-by-layer process mentioned above – vertical resolution dependent on the thickness of the layer stacked together.

Evanescent Light Stereolithography

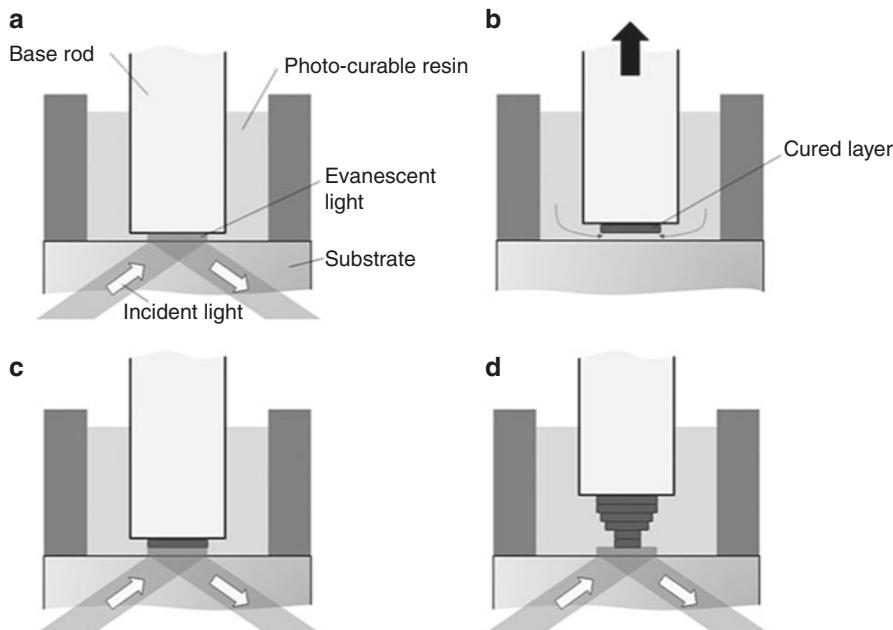
Evanescent light stereolithography, which uses an evanescent light for the photopolymerization of the liquid-state polymer differently from the conventional stereolithography technologies based on a propagating light, has also been developed [14] (see Fig. 7). When light travels across materials with different refractive indices at an incident angle greater than the so-called critical angle, above which total internal reflection of the light occurs at the interface, the evanescent light is generated within a localized area from the boundary of the substrate and the photo-curable polymer. As a result, the liquid-state polymer is hardened by the evanescent light in the sub-micrometer-thickness layer. By repeating the processes, which includes refilling the boundary with the liquid-state polymer and its photopolymerization by the evanescent light, 3D structures with sub-micrometer scale in vertical axis are formed. However, the lateral resolution of the objects formed by this technique is not very fine yet.



Stereolithography, Fig. 6 Schematic illustration of an image projection microstereolithography system using a dynamic image generated by a liquid crystal display (LCD)

Conclusion

SL has been a promising technology for the direct fabrication of complicated 3D structures in micro-/nanoscale. Although there has been a great deal of achievement in the stereolithographic technologies for several decades, there are still challenging issues, which should be explored for their practical uses: massive production, large-area fabrication, reliability of processes, and improvement of



Stereolithography, Fig. 7 Schematic illustration of evanescent light stereolithography processes. (a) First of all, an incident light exposes and cures the photo-curable polymer. (b) Next, the cured layer adhering on the base rod is lifted by raising the rod and filling the space with a new

layer of the liquid-state polymer. (c) Then a light beam with different specific 2D patterns exposes and cures the next layer. (d) Finally, the desired object can be fabricated by repeating this loop continuously

photopolymerizable materials. Nevertheless, the SL has been one of the most trustworthy technologies for the fabrication of 3D nano-/microstructures and would take a place as a powerful 3D fabrication technology, which is applicable to diverse research fields.

References

1. Kodama, H.: Automatic method for fabricating a three-dimensional plastic model with photo-hardening polymer. *Rev. Sci. Instrum.* **52**, 1770–1773 (1981)
2. Ikuta, K., Hirowatari, K.: Real three dimensional micro fabrication using stereo lithography and metal molding. In: Proceedings of IEEE Micro Electro Mechanical Systems, pp. 42–47. Fort Lauderdale (1993)
3. Ikuta, K., Ogata, T., Tsuboi M., Kojima, S.: Development of mass productive micro stereo lithography (Mass-IH process). In: Proceedings of IEEE Micro Electro Mechanical Systems, pp. 42–47. Fort Lauderdale (1993)
4. Ikuta, K., Maruo, S., Kojima, S.: New microstereolithography for freely movable 3D micro structures—super IH process with submicron resolution. In: Proceedings of IEEE Micro Electro Mechanical Systems, pp. 290–295. Heidelberg (1998)
5. Bertsch, A., Zissi, S., Jezequel, J.-Y., Corbel, S., Andre, J.C.: Microstereolithography using a liquid crystal display as dynamic mask-generator. *Microsyst. Technol.* **3**, 42–47 (1997)
6. Hadipoespiro, G., Yang, Y., Choi, H., Ning, G., Li, X.: Digital micromirror device based microstereolithography for micro structures of transparent photopolymer and nanocomposites. In: Proceedings of the Solid Freeform Fabrication Symposium, pp. 13–24. Austin (2003)
7. EnvisionTEC.: <http://www.envisiontec.de>. Accessed 16 Mar 2011
8. Maruo, S., Nakamura, O., Kawata, S.: Three-dimensional microfabrication with two-photon-absorbed photopolymerization. *Opt. Lett.* **22**, 132–134 (1997)
9. Cumpston, B.H., Ananthavel, S.P., Barlow, S., Dyer, D.L., Ehrlich, J.E., Erskine, L.L., Heikal, A.A., Kuebler, S.M., Lee, I.-Y.S., McCord-Maughon, D., Qin, J., Rockel, H., Rumi, M., Wu, X.-L., Marder, S.R., Perry, J.W.: Two-photon polymerization initiators for three-dimensional optical data storage and microfabrication. *Nature* **398**, 51–54 (1999)
10. Sun, H.-B., Kawakami, T., Xu, Y., Ye, J.-Y., Matuso, S., Misawa, H., Miwa, M., Kaneko, R.: Real three-dimensional microstructures fabricated by

- photopolymerization of resins through two-photon absorption. *Opt. Lett.* **25**, 1110–1112 (2000)
11. Kawata, S., Sun, H.-B., Tanaka, T., Takada, K.: Finer features for functional microdevices. *Nature* **412**, 697–698 (2001)
 12. Yang, D.-Y., Park, S.H., Lim, T.W., Kong, H.-J., Yi, S.W., Yang, H.K., Lee, K.-S.: Ultraprecise microreproduction of a three-dimensional artistic sculpture by multipath scanning method in two-photon photopolymerization. *Appl. Phys. Lett.* **90**, 013113 (2007)
 13. Park, S.-H., Yang, D.-Y., Lee, K.-S.: Two-photon stereolithography for realizing ultraprecise three-dimensional nano/microdevices. *Laser Photonics Rev.* **3**, 1–11 (2009)
 14. Kajihara, Y., Inazuki, Y., Takahashi, S., Takamasu, K.: Study of nano-stereolithography using evanescent light. In: Proceedings of the American Society for Precision Engineering (ASPE) Annual Meeting, pp. 149–152. Orlando (2004)

Stimuli-Responsive Drug Delivery Microchips

Jian Chen¹, Jason Li², Michael Chu²,
Claudia R. Gordijo², Yu Sun³ and Xiao Yu Wu²
¹State Key Laboratory of Transducer Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing, People's Republic of China
²Advanced Pharmaceutics and Drug Delivery Laboratory, Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON, Canada
³Department of Mechanical and Industrial Engineering and Institute of Biomaterials and Biomedical Engineering and Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada

Synonyms

Intelligent drug delivery microchips; MEMS-based drug delivery devices; Smart drug delivery microchips

Definition

Stimuli-responsive drug delivery microchips are MEMS-based “smart” drug delivery devices

composed of individually sealed drug reservoirs that can be opened selectively for complex drug release by various stimuli, targeting long-term implantation applications.

Overview

Overview of Working Mechanism

Advances in MEMS technology have enabled the precise fabrication of miniature biomedical devices with micrometer-sized features for implantable drug delivery. Drug delivery microchips contain small reservoirs that are loaded with drugs and separated from the outside environment by a drug release barrier. Examples of reported MEMS devices utilize various approaches including electrochemical dissolution, electrothermal activation, chemical degradation, or self-regulation to control temporal drug release or modulate the permeability of the drug release barrier for drug delivery on demand.

Advantages

These microreservoir devices are well suited for applications in chronotherapy due to their ability to achieve timed or regulated drug release with well-defined temporal profiles of drug depots and release mechanisms. In addition, microchip-based implantable drug delivery devices allow for localized delivery by direct placement of the device at the treatment site. Another benefit is that these devices contain no moving parts and are capable of delivering reservoir drugs in a solid, liquid, or gel formulation. The device design also protects the drug depot from the outside environment before release so that stability of the drug inside the reservoirs can be maintained more effectively in the controlled environment.

Potential Concerns

Implantable microchips require minor surgery for implantation and removal, and need to maintain their functions during the application. Therefore, these systems must be biologically inert to minimize inflammatory response and both chemically and physically stable to avoid premature device failure. While the implants need to be as

unobtrusive as possible, there is a fundamental limitation of device size imposed by the need for sufficient storage capacity for a chronic dosing regimen. Suitable candidate drugs will be potent and prepared in concentrate formulations and are stable for extended periods at body temperature in order to minimize implant size.

Methodology

Currently available drug delivery microchips can be divided into three main categories, based on drug release methods: active drug release (e.g., electrochemical dissolution and electrothermal activation approaches), passive drug release (e.g., chemical degradation), and self-regulated drug release (e.g., pH-responsive drug delivery).

Electrochemical Dissolution Approach

As shown in Fig. 1, the device is fabricated by the sequential processing of a silicon wafer using microelectronic processing techniques including UV photolithography, chemical vapor deposition, electron beam evaporation, and reactive ion etching. Samples of reservoirs are sealed at one end by a thin membrane of gold to serve as an anode in an electrochemical reaction. Electrodes are placed on the device to serve as a cathode.

When the microfabricated reservoir system is submerged in an electrolyte solution, ions form a soluble complex with the anode material in its ionic form. An applied electric potential oxidizes the anode membrane, forming a soluble complex with the electrolyte ions. The complex dissolves in the electrolyte, the membrane disappears, and the solution within the reservoir is released. The release time from each individual reservoir is determined by the time at which the reservoir's anode membrane is removed.

Release studies demonstrate that the activation of each reservoir can be controlled individually, creating a possibility for achieving many complex release patterns. Varying amounts of chemical substances in a solid, liquid, or gel form can be released into solution in a pulsatile or continuous manner, or a combination of both, either sequentially or simultaneously from a single device.

Such a device has additional potential advantages including small size, quick response times, and low power consumption. In addition, all chemical substances to be released are stored in the reservoirs of the microchip itself, creating a possibility for the future development of autonomous devices.

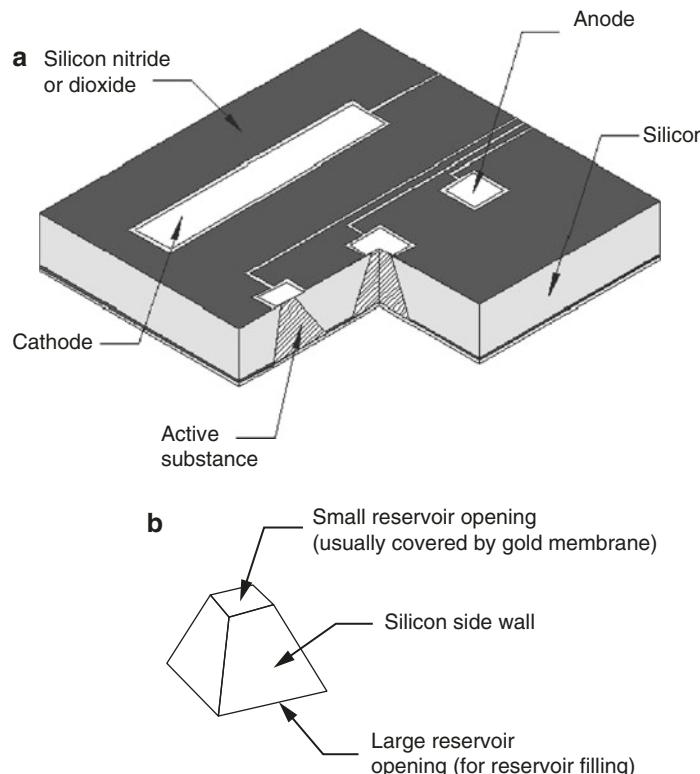
Electrothermal Activation Approach

The devices for this study were produced using standard microfabrication processes which contain an array of individually sealed and actuated reservoirs, each capped by a thin metal membrane comprised of either gold or multiple layers of titanium and platinum (see Fig. 2). The passage of a threshold level of electric current through the membrane causes it to disintegrate, thereby exposing the protected drugs of the reservoir to the surrounding environment. Compared to the electrochemical dissolution approach, electrothermal reservoir opening is more reliable and repeatable than the gradual opening achieved by the corrosion-based method.

In addition, reservoirs were aseptically sealed with spheres of indium-tin eutectic solder by thermocompression bonding. Filled and sealed microchips were electrically connected to the wireless communication hardware, power supply, and circuit boards of the in vivo implant, which were hermetically sealed inside a laser welded titanium case. Pulsatile release of a therapeutic polypeptide on demand, in response to telemetry between an external device controller and the implant, was demonstrated. Importantly, although a tissue capsule formed around the device (as expected), the capsule did not significantly affect the drug release profile during the 6-month implantation. This result addresses a major perceived risk of implanted drug delivery devices that a tissue barrier will compromise the effectiveness and control of release.

Chemical Degradation Approach

Biodegradable polymer version multireservoir drug delivery microchips have also been designed to investigate the feasibility of achieving multipulse drug release from a polymeric system over periods of several months without requiring



Stimuli-Responsive Drug Delivery Microchips,

Fig. 1 Schematic illustration of the first implantable drug delivery microchip via electrochemical dissolution approach. (a) Fabrication of these microchips began by depositing low stress, silicon-rich nitride on both sides of prime grade silicon wafers using a vertical tube reactor. The silicon nitride layer on one side of the wafer was patterned by photolithography and electron cyclotron resonance-enhanced reactive ion etching to give a square device containing square reservoirs. The silicon nitride served as an etch mask for potassium hydroxide solution, which anisotropically etched square pyramidal reservoirs (b) into the silicon along the (111) crystal planes until the silicon nitride film on the opposite side of the wafer was

reached. The newly fabricated silicon nitride membranes completely covered the *square* openings of the reservoir. Gold electrodes were deposited and patterned over the silicon nitride membranes by electron beam evaporation and liftoff. A layer of plasma-enhanced chemical vapor deposition silicon dioxide was deposited over the entire electrode-containing surface. The silicon dioxide located over portions of the anode, cathode, and bonding pads were etched with reactive ion etching to expose the underlying gold film; this technique was then used to remove the thin silicon nitride and chromium membranes located in the reservoir underneath the gold anode (Reproduced with permission from Ref. [1])

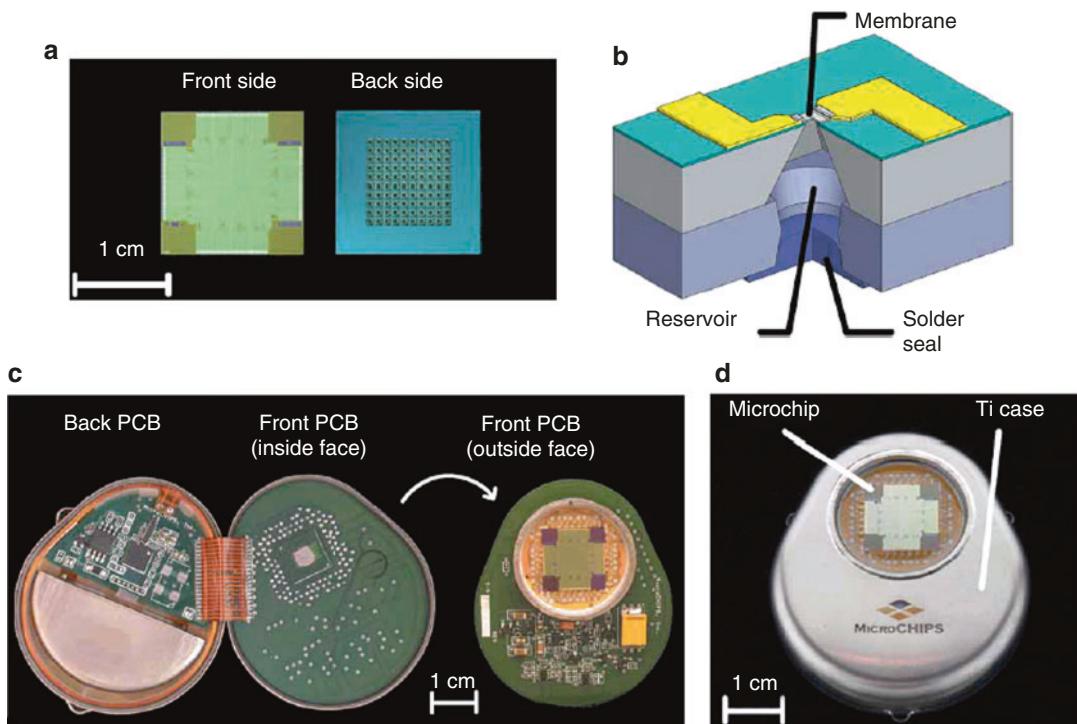
the application of a stimulus to trigger the drug release. A prototype device is shown in Fig. 3, from which drugs can be released at different times on the basis of the characteristics of the reservoir membranes that affect their degradation rate, such as the material used, the molecular mass, the composition, or the thickness.

Separation of the release formulation (the reservoir membranes) from the drug formulation (which is loaded into the reservoirs behind the membranes) might enable greater flexibility in

adapting this system for a desired application than is currently achievable with existing methods. An advantage of biodegradable polymeric microchips is the elimination of a requirement for a second surgery to remove the device. In addition, the lack of electronics reduces size restrictions in terms of device manufacture.

Self-regulated Approach

Compared to microchips with release mechanisms of electrochemical dissolution, electrothermal



Stimuli-Responsive Drug Delivery Microchips,

Fig. 2 Schematic illustration of the implantable drug delivery microdevice via electrothermal dissolution approach under remote control. (a) Front and back of the 100-reservoir microchip. (b) Representation of a single

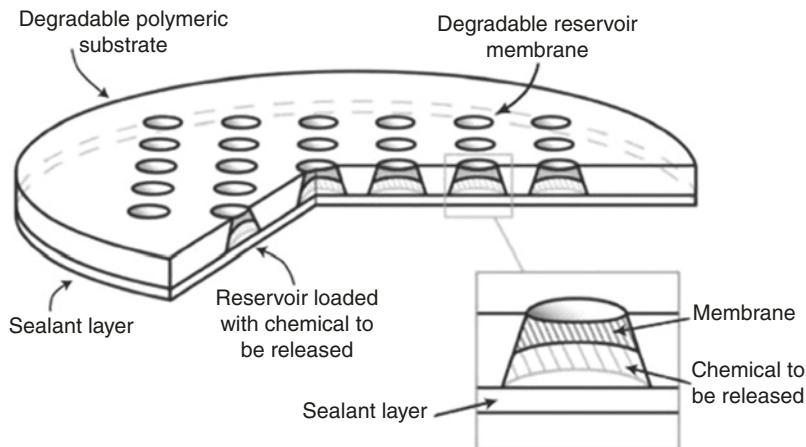
reservoir. (c) Electronic components on the printed circuit board (*PCB*) in the device package. (d) The assembled implantable device (Reproduced with permission from Ref. [2])

activation, and chemical degradation, which are not capable of regulating drug delivery rates in response to *in vivo* environmental conditions, a self-regulated microdevice enables drug release in response to local pH or glucose variations by integrating pH and/or glucose-responsive nanohydrogel embedded in composite membranes functioning as intelligent nanovalves [4–10].

As shown in Fig. 4, the patterned PDMS structure forms a drug reservoir and provides physical support for a thin nanohydrogel-embedded composite membrane. The hydrogel nanoparticles detect environmental pH changes and respond with corresponding volumetric swelling or shrinking, which regulates membrane permeability and thereby drug release rate. The polymeric microchips are monolithic without requiring peripheral control hardware or additional components for controlling drug release rates.

By adjusting nanoparticle percentages, membrane rigidity, drug reservoir shape and size, and drug loading volume and concentration, well-controlled drug release profiles in response to local pH changes can be achieved, functioning as a platform technology for intelligent drug delivery.

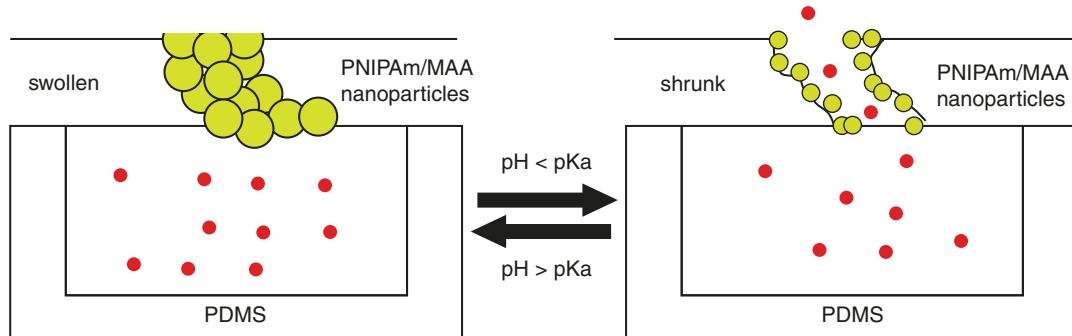
Recent studies have employed this approach for glucose-responsive insulin release. Glucose sensitivity combined with pH responsiveness is achieved with the use of glucose oxidase, an enzyme that catalyzes the oxidation of glucose to produce gluconic acid which in turn reduces local pH. *In vitro* insulin release profiles demonstrate rapid and repeatable device response to normal and hyperglycemic conditions [6, 8, 9]. Implantation of microdevices into diabetic rats resulted in rapid return to euglycemia following glucose challenges and successfully maintained normal blood glucose levels for several days [8, 9].



Stimuli-Responsive Drug Delivery Microchips,

Fig. 3 Schematic illustration of the first implantable drug delivery microdevice with biodegradable membranes. The main body of the device is composed of a reservoir-containing substrate that is fabricated from a degradable polymer. Truncated conical reservoirs in the substrate are

loaded with the chemical to be released and sealed with polymeric degradable reservoir membranes on one end and a sealant layer (polyester tape) on the opposite end. *Inset*, close-up of a reservoir, reservoir membrane, sealant layer, and chemical to be released (Reproduced with permission from Ref. [3])



Stimuli-Responsive Drug Delivery Microchips,

Fig. 4 Illustration of the mechanism for pH-responsive drug release out of the microdevice. *Left*: Nanoparticles are in the swollen state when the surrounding pH value is higher than pK_a (acid dissociation constant) of the

nano particles. *Right*: Nanoparticles are in the shrunk state when the surrounding pH value is lower than pK_a . Resulting volumetric swelling and shrinking of the nanoparticles control drug release rates [5, 7]

Key Research Findings

Device Dimension and Shape

For versatility and acceptability, overall device size and shape are major considerations. To be commercially attractive, the device should be small enough to be implanted subcutaneously in a doctor's office, with local anesthesia and a profile that is not obtrusive or apparent. The shape and materials of the device must allow easy

removal when the treatment is no longer needed or the device requires replacement.

MEMS-based implants must accommodate a sufficient quantity of drug doses in the device of minimum size, creating a design hurdle and constraint. This constraint requires the nondrug components to occupy minimal volume. Suitable candidate drugs should be potent, prepared in high-concentration formulations, and be stable for extended periods at the body temperature.

Materials

To minimize chronic inflammation and immune response, the tissue-contacting components must be constructed from biocompatible materials [11]. These materials must be stable in physiological fluids and tissues and nontoxic over the lifetime of the device. In addition, leachable from metal and polymeric materials should be minimal, nontoxic, and hypoallergenic. Silicon has traditionally been the most common material used for microdevice construction. More recent, device designs typically make use of biologically inert polymers such as polymethyl methacrylate (PMMA) and polydimethylsiloxane (PDMS) for device construction. These polymers have superior properties to silicon for BioMEMS applications with regard to cost and versatility of physical properties.

Inflammatory Response and Biocompatibility

Implantation of a foreign material within the body activates the host inflammatory response, which may influence device function and longevity through local production of reactive oxygen species, cell-mediated degradation, or formation of an impermeable fibrous capsule around the device [12]. The type and severity of the inflammatory response depend on the nature of the implant material, the geometry, surface topography, and various surface treatments of the device [10, 13, 14]. Thus, these parameters may be tuned to minimize the host inflammatory response and improve the overall biocompatibility of the device.

Biologically inert polymers can be used to mask non-biocompatible materials from the body in cases where their interaction with biological tissues is not required [15]. However, most

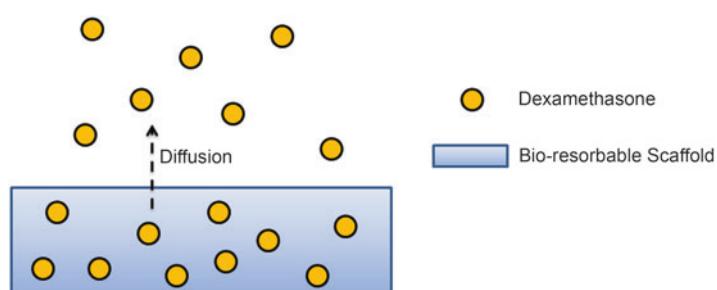
stimuli-responsive drug delivery systems feature non-biocompatible active components such as metallic or semiconducting sensing electrodes, stimuli-responsive polymers, or enzymes that must contact and interact with the host tissue or biochemicals for proper device functions. For these components, several strategies, namely, active and passive strategies, have been devised for mitigating the immune response while simultaneously allowing efficient diffusion of molecules to and from the device [16].

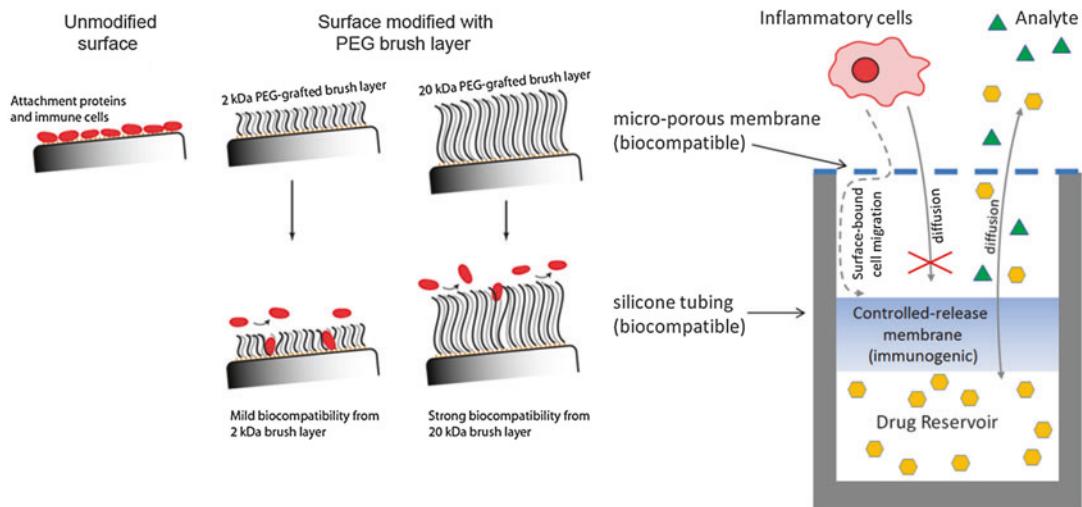
Active strategies for attenuating the inflammatory response involve the sustained local release of anti-inflammatory agents such as dexamethasone [17], nitric oxide [18], sirolimus [19], or paclitaxel [20] to suppress the normal wound healing process (Fig. 5). Care must be taken when employing these agents as they may cause adverse side effects or complications.

Passive strategies reduce the inflammatory response by minimizing serum protein adsorption and immune cell adhesion onto the implant surface. Steric hindrance is generated using non-fouling self-assembled monolayers, polymer brushes, or hydrogel coatings made from hydrophilic polymers such as poly(ethylene glycol) (Fig. 6a) [10, 21, 22] or surface-supported phospholipid bilayers [23]. In doing so, protein adsorption onto the implant surface is minimized, and thus, downstream recruitment of inflammatory cells is prevented. Alternatively, the inflammatory response can be minimized by restricting access of immune cells themselves to the implant surface by covering the implant with nano-porous size-exclusion membranes [24] or with the use of PEGylated microporous membranes to geometrically impede cell migration to the implant surface (Fig. 6b) [25].

Stimuli-Responsive Drug Delivery Microchips,

Fig. 5 Active strategy for mitigating the host inflammatory response involving the sustained release of anti-inflammatory agents from the implant surface





Stimuli-Responsive Drug Delivery Microchips,

Fig. 6 Representative passive strategies for mitigating the host inflammatory response following device implantation. (a) Use of non-fouling self-assembled monolayers, polymer brushes, or hydrogel coatings to minimize protein

adsorption onto implant surfaces (Reproduced with permission from Ref. [10]). (b) Use of microporous membranes and device geometry for hindering cell migration to the active implant surface [25]

Future Work

MEMS and miniaturization technologies have been used for producing novel stimuli-responsive drug delivery microchips that require minimal invasiveness of the medical procedures for implantation. Translation of this technology to the clinical setting will require improvements in implant reliability, safety, and lifetime. To achieve these goals, research efforts will focus on improving implant biocompatibility and development of stable and highly concentrated drug formulations.

Electrical systems will further require development of robust electrical components for selectively addressing individual reservoirs, biosensors for information acquisition, wireless communication hardware for remote control, and a long-term power source. These drug delivery implants will be tailored for personalized medicine and other unmet medical need [26].

References

- Santini Jr., J.T., Cima, M.J., Langer, R.: A controlled-release microchip. *Nature* **397**, 335–338 (1999)

- Prescott, J.H., Lipka, S., Baldwin, S., Sheppard Jr., N. F., Maloney, J.M., Coppeta, J., Yomtov, B., Staples, M.A., Santini Jr., J.T.: Chronic, programmed polypeptide delivery from an implanted, multireservoir microchip device. *Nat. Biotechnol.* **24**, 437–438 (2006)
- Grayson, A.C.R., Choi, I.S., Tyler, B.M., Wang, P.P., Brem, H., Cima, M.J., Langer, R.: Multi-pulse drug delivery from a resorbable polymeric microchip device. *Nat. Mater.* **2**, 767–772 (2003)
- Yam, F., Wu, X., Zhang, Q.: A novel composite membrane for temperature-and pH-responsive permeation. In: Park, K., Mrsny, R.J. (eds.) *Controlled Drug Delivery: Designing Technologies for the Future*. ACS Symposium Series, pp. 263–272. American Chemical Society, Washington, DC (2000)
- Zhang, K., Wu, X.Y.: Temperature and pH-responsive polymeric composite membranes for controlled delivery of proteins and peptides. *Biomaterials* **25**, 5281–5291 (2004)
- Zhang, K., Wu, X.Y.: Modulated insulin permeation across a glucose-sensitive polymeric composite membrane. *J. Control. Release* **80**, 169–178 (2002)
- Chen, J., Chu, M., Koulajian, K., Wu, X.Y., Giacca, A., Sun, Y.: A monolithic polymeric microdevice for pH-responsive drug delivery. *Biomed. Microdevices* **11**, 1251–1257 (2009)
- Gordijo, C.R., Shuhendler, A.J., Wu, X.Y.: Glucose-responsive bioinorganic nanohybrid membrane for self-regulated insulin release. *Adv. Funct. Mater.* **20**, 1404–1412 (2010)
- Gordijo, C.R., Koulajian, K., Shuhendler, A.J., Bonifacio, L.D., Huang, H.Y., Chiang, S., Ozin, G. A., Giacca, A., Wu, X.Y.: Nanotechnology-enabled

- closed loop insulin delivery device: in vitro and in vivo evaluation of glucose-regulated insulin release for diabetes control. *Adv. Funct. Mater.* **21**, 73–82 (2011)
10. Chu, M.K., Gordijo, C.R., Li, J., Abbasi, A.Z., Giacca, A., Plettenburg, O., Wu, X.Y.: In vivo performance and biocompatibility of a subcutaneous implant for real-time glucose-responsive insulin delivery. *Diabetes Technol. Ther.* **17**, 255–267 (2015)
 11. Nichols, S.P., Koh, A., Storm, W.L., Shin, J.H., Schoenfisch, M.H.: Biocompatible materials for continuous glucose monitoring devices. *Chem. Rev.* **113**, 2528–2549 (2013)
 12. Anderson, J.M., Rodriguez, A., Chang, D.T.: Foreign body reaction to biomaterials. *Semin. Immunol.* **20**, 86–100 (2008)
 13. Parker, J.A., Walboomers, X.F., Von den Hoff, J.W., Maltha, J.C., Jansen, J.A.: The effect of bone anchoring and micro-grooves on the soft tissue reaction to implants. *Biomaterials* **23**, 3887–3896 (2002)
 14. Moshayedi, P., Ng, G., Kwok, J.C.F., Yeo, G.S.H., Bryant, C.E., Fawcett, J.W., Franze, K., Guck, J.: The relationship between glial cell mechanosensitivity and foreign body reactions in the central nervous system. *Biomaterials* **35**, 3919–3925 (2014)
 15. Muskovich, M., Bettinger, C.J.: Biomaterials-based electronics: polymers and interfaces for biology and medicine. *Adv. Healthcare Mater.* **1**, 248–266 (2012)
 16. Bridges, A.W., García, A.J.: Anti-inflammatory polymeric coatings for implantable biomaterials and devices. *J. Diabetes Sci. Technol.* **2**, 984–994 (2008)
 17. Blanco, E., Weinberg, B.D., Stowe, N.T., Anderson, J. M., Gao, J.: Local release of dexamethasone from polymer millirods effectively prevents fibrosis after radiofrequency ablation. *J. Biomed. Mater. Res. A* **76**, 174–182 (2006)
 18. Hetrick, E.M., Prichard, H.L., Klitzman, B., Schoenfisch, M.H.: Reduced foreign body response at nitric oxide-releasing subcutaneous implants. *Biomaterials* **28**, 4571–4580 (2007)
 19. Choi, J., Jang, B.N., Park, B.J., Joung, Y.K., Han, D.K.: Effect of solvent on drug release and a spray-coated matrix of a sirolimus-eluting stent coated with poly(lactic-co-glycolic acid). *Langmuir* **30**, 10098–10106 (2014)
 20. Ren, K., Zhang, M., He, J., Wu, Y., Ni, P.: Preparation of polymeric prodrug paclitaxel-poly(lactic acid)-b-polylisobutylene and its application in coatings of drug eluting stent. *ACS Appl. Mater. Interfaces* (2015). doi:10.1021/acsami.5b01410
 21. Lokanathan, A.R., Zhang, S., Regina, V.R., Cole, M.A., Ogaki, R., Dong, M., Besenbacher, F., Meyer, R.L., Kingshott, P.: Mixed poly (ethylene glycol) and oligo (ethylene glycol) layers on gold as nonfouling surfaces created by backfilling. *Biointerphases* **6**, 180–188 (2011)
 22. Li, P., Poon, Y.F., Li, W., Zhu, H.-Y., Yeap, S.H., Cao, Y., Qi, X., Zhou, C., Lamrani, M., Beuerman, R.W., Kang, E.-T., Mu, Y., Li, C.M., Chang, M.W., Jan Leong, S.S., Chan-Park, M.B.: A polycationic antimicrobial and biocompatible hydrogel with microbe membrane suctioning ability. *Nat. Mater.* **10**, 149–156 (2011)
 23. Glasmöstar, K., Larsson, C., Höök, F., Kasemo, B.: Protein adsorption on supported phospholipid bilayers. *J. Colloid Interface Sci.* **246**, 40–47 (2002)
 24. de Vos, P., Faas, M.M., Strand, B., Calafiore, R.: Alginate-based microcapsules for immunoisolation of pancreatic islets. *Biomaterials* **27**, 5603–5617 (2006)
 25. Li, J., Chu, M.K., Gordijo, C.R., Abbasi, A.Z., Chen, K., Adissu, H.A., Lohn, M., Giacca, A., Plettenburg, O., Wu, X.Y.: Microfabricated microporous membranes reduce the host immune response and prolong the functional lifetime of a closed-loop insulin delivery implant in a type 1 diabetic rat model. *Biomaterials* **47**, 51–61 (2015)
 26. Chertok, B., Webber, M.J., Succi, M.D., Langer, R.S.: Drug delivery interfaces in the 21st century: From science fiction ideas to viable technologies. *Mol. Pharm.* **10**, 7 (2013).
-
- ## Stimulus-Responsive Polymeric Hydrogels
- ▶ Smart Hydrogels
-
- ## Stochastic Assembly
- ▶ Self-Assembly for Heterogeneous Integration of Microsystems
-
- ## Strain Gradient Plasticity Theory
- ▶ Plasticity Theory at Small Scales
-
- ## Structural Color in Animals
- Mathias Kolle¹ and Ullrich Steiner²
- ¹Harvard School of Engineering and Applied Sciences, Cambridge, MA, USA
- ²Department of Physics, Cavendish Laboratories, University of Cambridge, Cambridge, UK
-
- ## Synonyms
- Animal coloration; Biological photonic structures; Biological structural color; Bio-optics; Bio-photonics

Definition

Intense and bright colors result from the interaction of light with periodic micro- and nanostructures that cause color by interference, coherent scattering, or diffraction. These colors are termed structural colors, and structures that cause color by modulation of light are called photonic structures. Photonic structures are usually composed of regular lattices with periodicities on the order of the wavelength of light. Various organisms in nature are known to use intriguingly diverse photonic structures.

Introduction

Structural Colors and Photonic Structures

Structural colors in the animal kingdom have attracted increasing research interest in recent years. Biological organisms offer an enormous variety of periodic micro- and nanostructures that by specific interaction with light provide distinct coloration. This sometimes dynamic reflectivity is tailor-made for the organism's purpose within its natural illumination environment. Intriguing photonic structures have been identified on the wing cases and armors of beetles, the scales of butterflies, the feathers of birds, in the shells of marine animals, or even within the skin of mammals. Nature offers a huge choice of blueprints for novel artificial optical materials and photonic structures. The strongest color contrasts are achieved by a combination of different physical effects, including multilayer interference, diffraction, coherent scattering, and spatially confined absorption [1–3]. A common design concept in natural photonic systems is the complex interplay of structural regularity on the length scale of the wavelength of visible light combined with structural disorder on a larger scale [4]. In many cases, the complex interaction of these hierarchical structures with incident light lead to an outstanding, dynamic coloration, bright reflectivity that is perceivable in a wide angular range, brilliant whiteness, or enhanced transmission [5, 6]. While photonic structures can be made entirely from transparent materials, the incorporation of absorbing

pigments deposited under or incorporated into a photonic structure is frequently used in nature to prevent spurious reflections. This improves the contrast leading to an enhancement of the color perceived from the photonic system [7]. Biological micro- and nanostructures that cause structural color are very diverse and often show periodicities on several length scales. This makes the optical characterization of a biological photonic system and the determination of the optical properties of its constituent materials (e.g., refractive index) very challenging. A set of useful techniques for the determination of the complex refractive indices of materials in biological photonic structures has recently been reported and successful attempts have been made to determine the complex refractive index of the organic cuticle material in the scales of butterflies and the armor of beetles [8].

Occurrence and Purpose of Structural Colors in Nature

Structural colors can be found in the feathers and skin of various birds [9–11], in the shells, spines, and scales of marine animals [12, 13], and in the skin of some mammalian species [14]. They are probably most abundant in species of the insect orders *Lepidoptera* and *Choleoptera* that comprise butterflies, moths, and beetles [15–19]. The different purposes of structural colors are as diverse as the organisms that use them [20]. Intense colors with stark contrast to their environment can serve in interspecies interaction including agonistic displays to confuse or scare away potential predators, while structural colors are also applied to induce a cryptic coloration for camouflage. For other animals, they are playing an import role for intraspecies communication such as competition between males of the same species for territory and/or females. Structural colors often provide sexual dichroism between males and females and are believed to have a function in sexual selection. They might also play a role in temperature regulation. While the physics and functioning of photonic structures found in many different organisms are mostly understood, it is frequently very challenging to clearly identify the specific benefit for the animal, which often remains mysterious.

The Physical Effects Underlying Structural Colors

Bright, pure, and intense colors arise from strong reflectivity in narrow spectral bands caused by highly ordered structures including multilayers, surface diffraction gratings, and photonic crystals. While many biological photonic structures are based on multilayer assemblies and two- or three-dimensional photonic crystals [1], diffraction gratings are rare in nature, possibly because they do not display a specific color but give rise to a range of colors depending on light incidence and observation direction [4]. The contrast of structural colors in their environment is often increased by the placement of absorbing elements spatially under or around the photonic structure. Intense blacks are achieved in nature by structurally assisted pigmentation as in the case of the butterfly *Papilio ulysses* [7]. In some avian species the pigments are incorporated directly into the photonic structure [9]. Brilliant white is achieved by multiple, incoherent scattering caused by highly disordered structures of random size, aperiodically arranged on the length scale of the wavelength of visible light [6]. Intense, spatially homogenous, angle-independent colors arise from coherent scattering caused by structures of well-defined size comparable to the wavelength of light, such as air pores of a narrow size distribution dispersed randomly in the volume of a material. This, for example, gives rise to the strong blue of the feathers of several different bird species [11]. High transparency is achieved by graded refractive index surfaces, involving arrays of conical protrusions found in moth eyes [5].

In the following, the structural designs, which give rise to strong coloration, ultrahigh blackness, or brilliant whites are discussed with emphasis on the interplay of different structural, hierarchically assembled elements. Specific natural organisms that apply these particular structural combinations are presented in an exemplary manner, without recounting all the animals known to apply the same or very similar concepts. First, static systems that induce structural colors based on “simple” multilayer elements are discussed before progressing to more and more sophisticated

two- and three-dimensional structural arrangements. Attention is paid to combinations of different photonic elements and to the interplay of order and disorder in hierarchical structures on the nano- and microscale, a powerful concept occurring in natural structural colors.

Static Structural Colors in Animals

Multilayer Structures of Varying Complexity in Natural Photonic Systems

Thin film interference and multilayer interference are among the most common phenomena inducing structural color in animals. Light reflected from periodic stacks of multiple planar, optically distinct, transparent layers (often called Bragg mirrors) is colored, provided the thickness of the individual layers is comparable to the wavelength of light in the visible spectrum. The color, its intensity, and purity depend on the refractive index contrast of the multilayer materials, the individual layer thicknesses, and the number of layers in the stack [2]. Furthermore, the perceived color strongly varies with the angle of light incidence and the observation direction. The peak wavelength λ_{\max} of light reflected from a multilayer in air composed of two materials with refractive index n_1, n_2 and film thicknesses d_1, d_2 upon incidence at an angle θ is given by the relation

$$m \cdot \lambda = 2 \cdot (n_1 d_1 \cos \theta_1 + n_2 d_2 \cos \theta_2), \quad (1)$$

where m is a positive integer and the angles θ_1, θ_2 of the light path in the materials are given by Snell's law $n_{\text{air}} \sin \theta = n_1 \sin \theta_1 = n_2 \sin \theta_2$. In this case the rays reflected from the family of $n_1 - n_2$ -interfaces interfere constructively with each other, if the light has the wavelength λ_{\max} . The same holds for light rays reflected from the $n_2 - n_1$ -interfaces in the stack. If a second relation given by

$$\begin{aligned} \left(m + \frac{1}{2} \right) \cdot \lambda &= 2n_1 d_1 \cos \theta_1 \\ &= 2n_2 d_2 \cos \theta_2 \end{aligned} \quad (2)$$

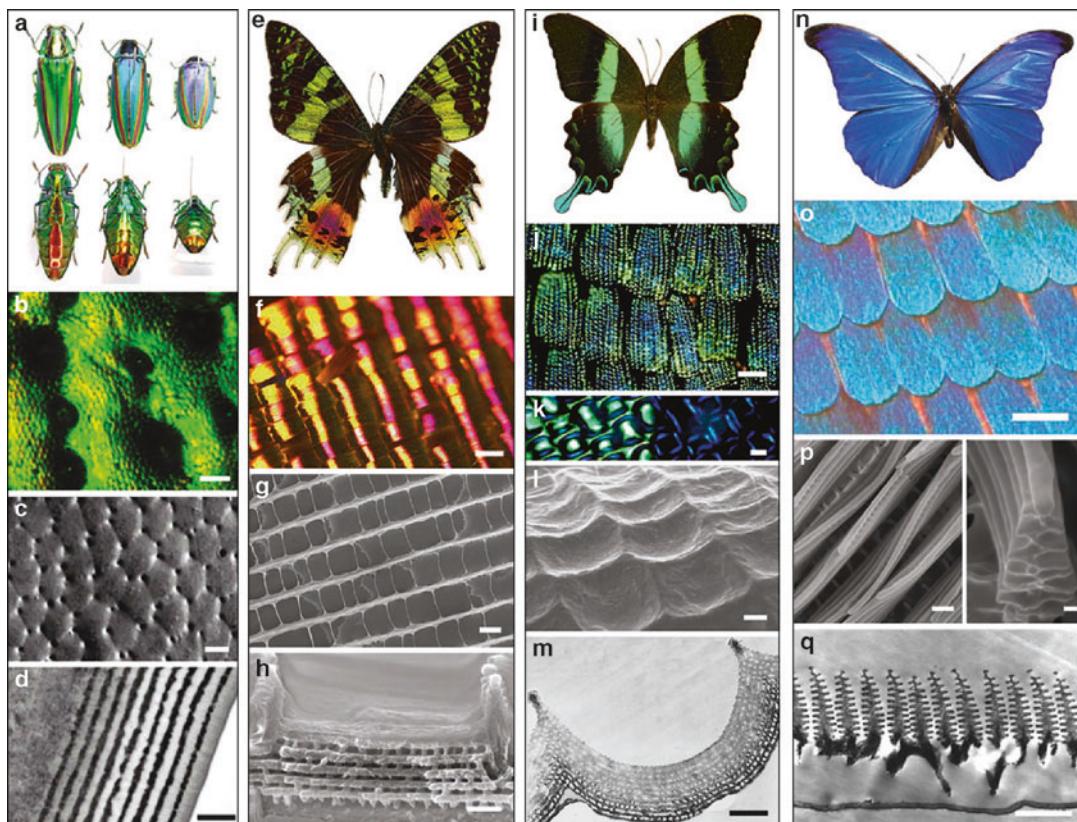
is also satisfied, the multilayer stack is referred to as an ideal multilayer. This relation signifies the fact that in an ideal multilayer the light rays reflected from the $n_1 - n_2$ -interfaces interfere constructively with the rays reflected from the $n_2 - n_1$ -interfaces, which is not the case for a “nonideal” multilayer. The reflectance of a nonideal multilayer is therefore lower compared to an ideal multilayer made of the same materials, with an equal overall layer number and equal periodicity $d_1 + d_2$. Most multilayer arrangements found in nature are nonideal in the sense of this classification. In some cases this may even be beneficial for the organism since nonideal multilayers have a narrower bandwidth compared to ideal multilayers, which increases the purity of the reflected color.

The wood-boring Japanese jewel beetle *Chrysochroa fulgidissima* found in the forests of Japan during summer displays a vivid ventral and dorsal coloration (Fig. 1a). This color originates from interference of light in a multilayer stack of about 20 alternating layers (Fig. 1d) with refractive indices of 1.5 and 1.7 incorporated in the epicuticle of the beetle’s wing cases [4]. The perceived coloration strongly depends on the orientation of the beetle with respect to the incident light and the observation direction, as expected for a multilayer structure (Fig. 1a). Nevertheless, the beetle’s surface does not act like a colored mirror, as expected from a perfectly flat multilayer arrangement. Minute hexagonally arranged holes and indentations are distributed in the multilayer on the 10 μm length scale (Fig. 1c) and the areas bordered by the micro-holes vary in their inclination. On the 100 μm length scale, the surface shows a strongly irregular corrugation (Fig. 1b). This structural irregularity leads to the scattering of light reflected from the multilayer, giving the beetle its characteristic color in a larger angular range around the specular reflection direction thus increasing its visibility.

The scales of the Madagascan moth *Chrysiridia rhipheus* (also called sunset moth because of the two bright yellow-orange-violet spots on the lower wing pair) display intense, beautiful colors ranging from green to red, similar to the jewel beetle’s hues (Fig. 1e, f).

The different, similarly bright colors result from a multilayer incorporated into the body of the moth’s wing scales (Fig. 1h). The multilayer is made from layers of one single material (cuticle) spaced by perpendicular struts, effectively creating a controlled air spacing between the cuticle layers. Including air as the second component in the multilayer stack leads to the maximization of the refractive index contrast, providing intense reflection with a much smaller number of layers compared to the Japanese jewel beetle, where the refractive index contrast is much lower. The air spacing varies in the differently colored regions of the moth’s wing with the smallest spacing found in the green areas and the largest spacing seen in the red wing spots. The multilayered scales show a strong curvature along their long axis. This curvature superimposed onto the photonic multilayer structure leads to a high visibility of the colors from various directions and also introduces multiple reflection effects between adjacent scales, leading to an increased purity of the reflected color and to interesting polarization effects [21]. Overall, caused by the interplay of the highly reflective air-cuticle multilayer and the pronounced scale curvature, the moth wing displays bright colors with a velvety, textile-like shimmer.

Air-cuticle multilayers of even more complex shapes can be found on the wing scales of butterflies of the genus *Papilio*. The South-East Asian Emerald Swallowtail *Papilio palinurus* and the Green Swallowtail *Papilio blumei* (Fig. 1i) display bright green spots on their wings resulting from a concavely shaped multilayer architecture on the individual scales (Fig. 1l, m). The vivid green originates from a superposition of blue and yellow reflected from the edges and the centers of the multilayer concavities, respectively (Fig. 1j, k) [22]. Different colors result from distinct regions of the concavely shaped multilayer due to the spatially varying angle of light incidence on the edges and in the center of the concavities. Light reflected from the edges can undergo multiple reflections within a single concavity, inducing a change in polarization. The butterfly *Papilio ulysses* displays concavely shaped multilayer structures on its scales that have less curvature and smaller cuticle and air gap thicknesses, resulting



Structural Color in Animals, Fig. 1 Natural photonic systems based on multilayer arrangements. (a) The wing cases and the ventral side of the Japanese Jewel beetle *Chrysochroa fulgidissima* display bright and iridescent colors that show a strong angular color variation (b, c). The surface of the wing cases shows irregular corrugations on the 100 μm scale (b, scale bar ~100 μm) and more regular indentations on the 10 μm scale (c, scale bar 10 μm). (d) Cross-sectional transmission electron micrographs reveal the stack of alternating high and low refractive index layers that causes multilayer interference, resulting in the beetle's iridescent color, scale bar 400 nm (Pictures (a–d) reproduced with permission of S. Kinoshita and IOP Publishing © 2008 Kinoshita et al. [1]). (e) The Madagascan moth, *Chrysiridia rhipheus*, displays a range of bright shimmering colors on its wings. (f) Different wing regions are covered by patterns of colorful scales reflecting green, yellow, red, or violet light. The scales are highly curved so that the observer only perceives light reflected from part of the scales, which creates the impression of texture caused by the juxtaposition of bright and dark regions, scale bar 200 μm . (g) The scales are spanned by parallel micro-ribbed ridges, scale bar 2 μm . (h) Cuticle layers of well-defined thickness that are spaced by small struts extend between and under the ridges, forming a regular multilayer arrangement with air as the

low refractive index material, scale bar 500 nm. (i) The butterfly *Papilio blumei* displays lucid green stripes on its upper and lower wing pairs. (j) At higher magnification, the green scales show regions of distinct blue or yellow color, scale bar 100 μm . (k) The yellow color from the centers of concavely shaped surface corrugations disappears when the scales are imaged between crossed polarizers while the blue from the edges of the concave shapes persists due to polarization rotation upon reflection, scale bar 5 μm . (l) Concave surface corrugations on the scales, scale bar 2 μm . (m) A cross-section through one of the concavities reveals cuticle layers of well-defined thickness and regular spacing forming a multilayer reflector, scale bar 1 μm (Figure (m) reproduced with permission of P. Vukusic and Springer Science + Business Media © 2009 Vukusic [19]). (n) The South American butterfly *Morpho rhetenor*. (o) Bright blue scales cover the wing membrane like tiles on a roof top, scale bar 100 μm . (p) Top and cross-sectional view of the ridges running along each scale, scale bars 500 and 100 nm. (q) A cross-section through a scale exposes the intricate design of the ridges, scale bar 1 μm . The horizontal extensions on each ridge act as a multilayer reflector tuned for the blue spectral range with air as the low refractive index medium (Figures (o and q) reproduced with permission of P. Vukusic and Springer Science + Business Media © 2009 Vukusic [19])

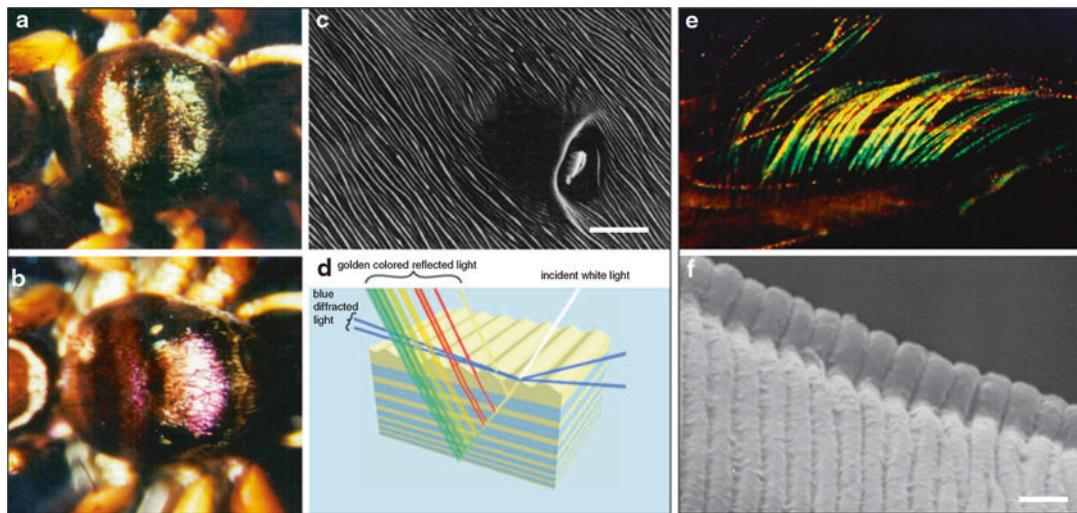
in a lucid blue color. While in the case of *Papilio palinurus* and *Papilio blumei* the higher local concavity wall curvature results in the juxtaposition of yellow and blue reflected from different regions forming green (a color that might have a purpose in camouflage), the shallower concavities of *Papilio ulysses* do not produce the same effect but rather diffuse the reflected light into a wide angular range for higher visibility [19].

The possibly most sophisticated and intensely studied multilayer structure giving rise to bright animal coloration is found on the wing scales of butterflies belonging to the genus *Morpho*. These butterflies that are native to South and Central America are known for their intense, widely visible coloration. Among them, *Morpho rhetenor* is the most striking example of vivid bright blue structural color (Fig. 1n, o). The intriguing hue of its wings results from the interference of light caused by ridge structures (Fig. 1p) on the wing scales that have a Christmas-tree like cross-section (Fig. 1p, q). The horizontal periodic protrusions of the ridges cause “quasi”-multilayer interference [4, 5]. The term “quasi” is used in this context, because light not only interferes when reflected from protrusions of the same ridge but also when reflected from protrusions on different adjacent ridges. As for the moth *Chrysiridia rhipheus* and the butterflies of the genus *Papilio*, the incorporation of air layers into the multilayer structure leads to a large refractive index contrast and consequently to a bright reflection in a wide spectral band. The ridges are spaced by $\lesssim 1 \mu\text{m}$ on average, which in the past lead to the assumption that they also act as a diffraction grating. However, local variation between the ridges in height and distance and in the orientation of entire scales confines the diffraction to individual ridges and inhibits the manifestation of a pronounced diffraction grating effect in reflection [4]. Nevertheless, this randomization seems to have a beneficial side effect: despite its origin from multilayer reflection, the blue color of the *Morpho rhetenor* is very illumination and observation angle insensitive and perceivable from all points in space above the wing plane, only changing to violet at very high angles of light incidence or observation.

Diffractive Elements

Diffraction from surface gratings is believed to be rare in nature. This might be due to the fact that diffraction gratings with periodicities of 400 nm–2 μm need long-range order on the 60 μm scale (the spatial coherence of sun light) in order to be efficient, a criterium that is hard to meet in the natural context where irregularity is usually predominant at this length scale. During the past decade some animals, mostly invertebrates, have been shown to employ diffraction gratings, however [12]. As opposed to multilayer reflectors, which reflect light in a relatively narrow spectral band, diffraction gratings split incident light into its spectral components and redirect each color into a different direction. Consequently, diffraction gratings are less suitable for providing spectrally well-defined colors in a wide angular range (which is better achieved with a multilayer reflector combined with some irregularity on the micron-scale). While diffraction gratings made from transparent materials efficiently create vivid colors in transmission, they have to be made from (or backed by) reflective materials to produce strong colors in reflection, which seems to be less efficient for the organism.

Periodic grating-like surface structures on an underlying chirped broadband multilayer reflector have been found on the torso of spiders [23]. In a chirped multilayer the thicknesses of the individual layers decrease or increase gradually from the top to the bottom of the stack. Constructive interference occurs for the reflection of light of a particular color/wavelength range at a specific depth in the stack (Fig. 2d) where the layer thicknesses fulfill the condition discussed above (Eq. 1) thereby leading to a broadband reflection usually resulting in a silver or golden color. In the case of the spider *Cosmophasis thalassina* (Fig. 2a, b), the striations (Fig. 2c) mainly serve the purpose to disperse blue light before it interacts with the underlying reflector (Fig. 2d), thereby biasing the color of the spider’s body toward a golden appearance, instead of the more silvery shine which it would have without the striations. Fourier analysis shows that there is no pronounced long-range order in these striations and



Structural Color in Animals, Fig. 2 Natural photonic systems based on diffraction. (a) The torso of the spider *Cosmophasis thalassina* shows two broad metallic golden-greenish stripes in air. (b) The color changes to purple-silvery when immersed in water. (c) The structure on the spider torso that causes these colors consists of a combination of surface striations superimposed on a chirped multilayer reflector (not shown here), scale bar 5 μm . (d) A schematic of the interaction of light with the spider's photonic structure. Blue light is very efficiently scattered by the striations while light of higher wavelength interferes

with the multilayer stack. Red is specularly reflected in the top section of the chirped mirror where the layers are thicker while green is reflected by the thinner layers further down the stack. (e) The hairs of the first antenna of the seed shrimp *Azygocypriidina lowryi* show strong iridescence caused by grating diffraction. (f) The diffraction grating on a single *setule* (hair) of the antenna, scale bar 1 μm (Images (d) modified) reproduced with permission of A. R. Parker, IOP Publishing © 2003 Parker and Hegedus [23] and The Royal Society © 2005 Parker [13])

consequently the striations do not create strong grating diffraction. The striations are likely to disperse light reflected from the multilayer reflector into a wider angular range. This represents a good example of the interplay of different structures with varying length scales to achieve a specific optical response.

The use of periodic diffraction gratings as surface structures seems particularly beneficial on narrow cylindrical geometries where the implementation of multilayers might be impractical or impossible. The thin hairs (*setae*) of some Crustacean species including the antenna of the male seed shrimp, *Azygocypriidina lowryi* (Fig. 2e), display vibrant colors resulting from periodic surface structures [12, 13]. In the antenna of *Azygocypriidina lowryi* the grating is formed by regular undulations of the hair thickness with a periodicity of 600–700 nm (Fig. 2f).

A curious structure causing diffraction with a reverse angular color sequence was recently

found on the wings of the male butterfly *Pierella luna* [24]. The ends of the scales in the central portions of this butterfly's forewings are curled upward, enabling regularly arranged, on average 440 nm spaced cross-ribs to act as an upward-directed diffraction grating in transmission, resulting in this intriguing phenomenon of "inverse" diffraction. As opposed to a conventional diffraction grating, these structures make the central forewing regions appear red for small angles of observation (measured from the surface normal), changing to yellow, green, and finally blue as the observation angle is increased.

Two-Dimensional Photonic Crystals

Two-dimensional photonic crystals that cause structural colors have been found in marine animals [13, 25, 26], birds [9, 10], and mammals [14]. This part of the review focuses on two-dimensional photonic systems found in nature that induce the striking color in some

animals. The *setae* of the polychaete worm *Pherusa* sp. (Fig. 3a, b) show a vivid play of colors originating from the periodic two-dimensional hexagonal arrangement of cylindrical channels (Fig. 3c, d) with a well-defined lattice constant within a single *seta* [26]. This system is strikingly similar to artificial photonic crystal fibers. A very similar arrangement was found earlier in the spines of the sea mouse *Aphrodisia* sp. [25] (Fig. 3e–h).

The feathers of kingfishers, peacocks, ducks, pigeons, and trogons among various other birds display strong blue and green colors that reportedly have attracted the interest of scientists for more than 300 years [1, 4, 18]. The bright blues and greens result from well-ordered melanin granules in a dense keratin matrix [9] or a spongy keratin network with regular air pores [11] of varying dimensions and structural complexity within the birds' feather barbules. These structures have been classified and investigated in great detail in the last three decades [9]. The order of the layers of the solid or hollow melanin granules or the pores in the keratin matrix varies from species to species. Consequently, the physical effect causing the bright color was repeatedly identified as incoherent scattering for the less ordered randomly distributed structural elements with well-defined sizes or as coherent scattering (for instance, multilayer interference) in structures of quasi-ordered arrangements with higher spatial order. In many bird species not only the feather barbules but also the skin on different body parts show strong color. These colors are based on coherent scattering caused by two-dimensional photonic crystal structures, consisting of regular arrays of collagen fibers in the dermis of the birds (Fig. 3i–l). The dimensions of the collagen fibers are very well defined with a narrow distribution in fiber diameter for each species, while the extent of spatial order varies from species to species (compare Fig. 3j, l).

Three-Dimensional Photonic Crystals

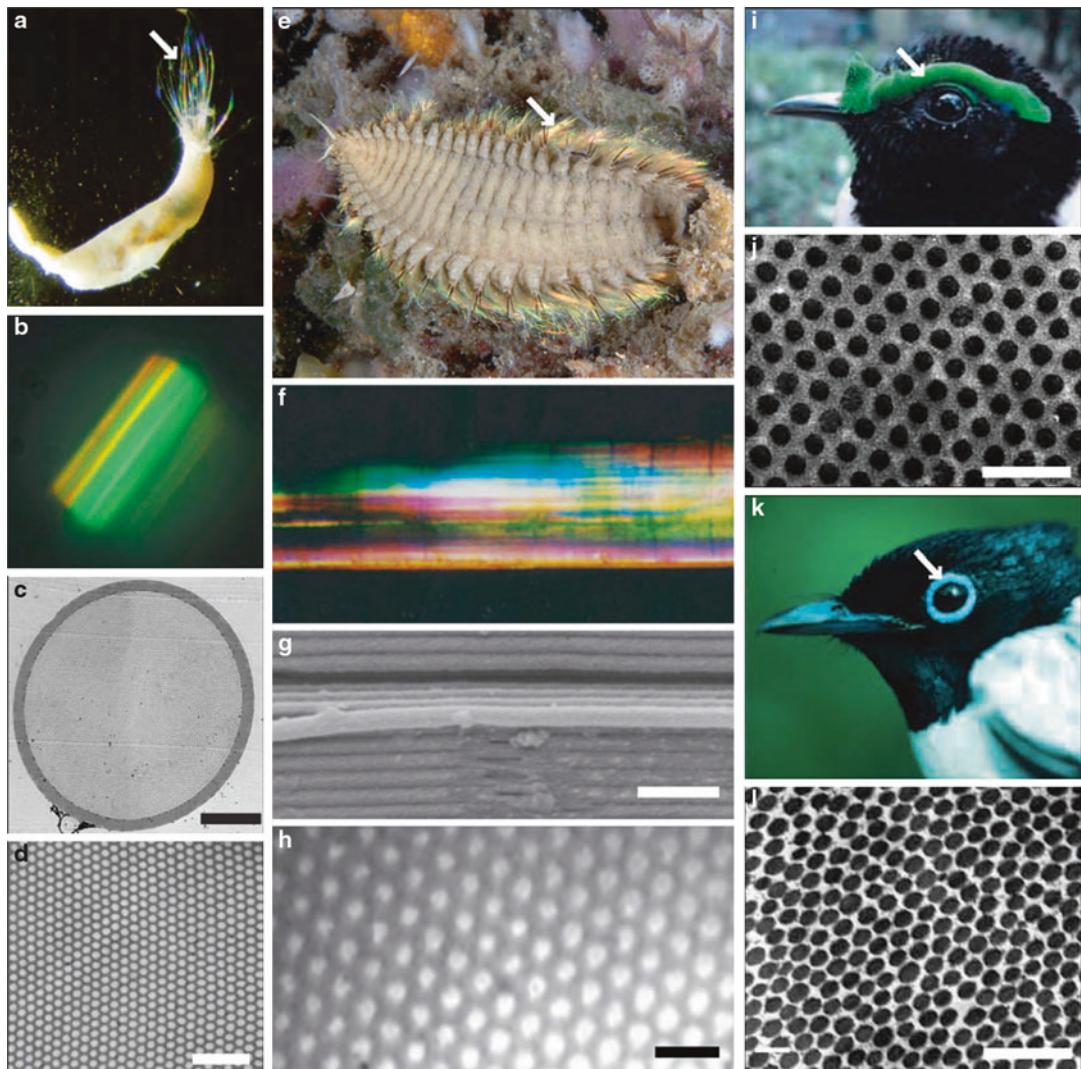
The first discovered biological three-dimensional photonic crystal is a solid, hexagonally close-packed array of transparent spheres of 250 nm diameter in a matrix of hydrated chitinous

material found inside the scales of the Australian weevil *Pachyrhynchus argus* [28] (Fig. 4a–c). Similar to the ordered arrangement of silica spheres in precious opal, this structure gives the beetle a metallic coloration visible in any direction. Recently, a similar structure was also observed in the Asian longhorn beetle *Pseudomyagrus waterhousei* (Fig. 4d–f).

Inverse opal photonic structures have been identified in various butterfly and beetle species. The beetle *Pachyrhynchus congestus pavonius* [30] displays vivid patches of orange color on its body (Fig. 4g, h) that originate from the interference of light within the beetle scales, which consist of a multilayered cortex surrounding an inverse opal photonic crystal (Fig. 4i). This is an example of the superposition of different structural geometries, in this case a multilayer mirror and a three-dimensional photonic crystal. The light reflected from a regular multilayer mirror results in an iridescent color while the interference of light with a polycrystalline three-dimensional photonic crystal usually leads to a uniform color over wide angular ranges. However, the precise influence of each structural component on the overall appearance of this particular beetle, *Pachyrhynchus congestus pavonius*, remains to be investigated in detail. The wings of the butterfly *Parides sesostris* display patches of bright green angle-independent color (Fig. 4j, k), resulting from the interaction of light with a polycrystalline, three-dimensional photonic crystal structure that is buried under the superficial ridging in the body of the butterfly's wing scales (Fig. 4l, m). The individual crystallites were shown to have cubic symmetry and recent research suggests that they consist in fact of a bi-continuous regular gyroid network [31]. The different orientations of the photonic crystal domains and the light scattering induced by the superposed ridge structure ensure that the reflected color is independent of observation angle.

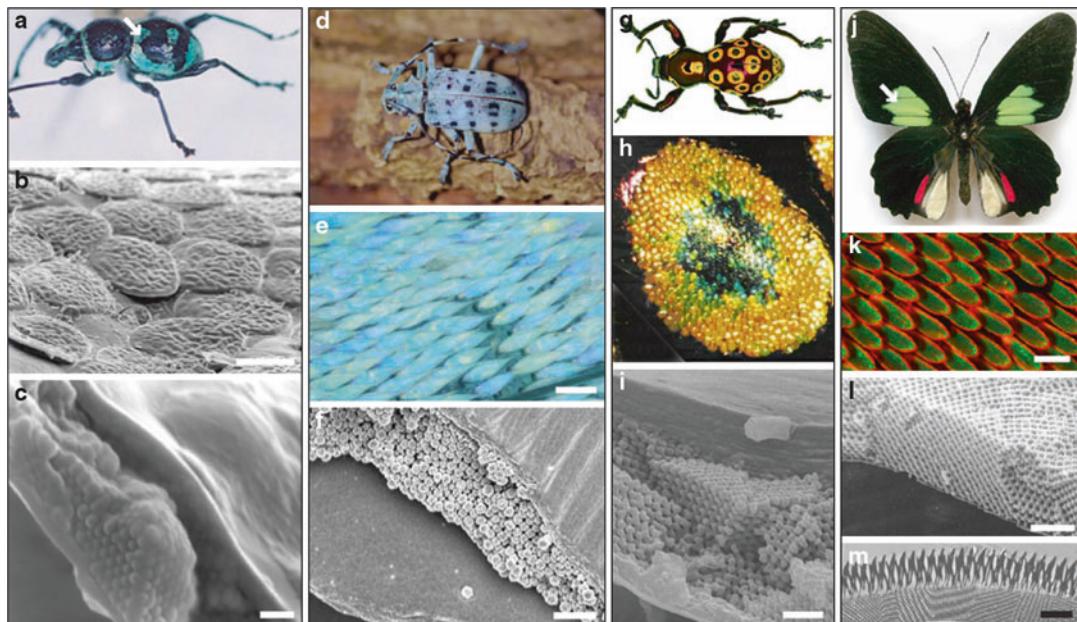
Structurally Assisted Blackness and Brilliant Whiteness

The natural photonic structures presented so far serve the organisms to display a distinct coloration. The contrast of a color is often enhanced in nature by placing it against a black background



Structural Color in Animals, Fig. 3 Structural colors based on two-dimensional photonic crystals. (a) The *setae* of the polychaete worm *Pherusa* sp. show a remarkable structural color (white arrow). (b) Micrograph of a *seta* of the worm. (c) A cross-sectional transmission electron micrograph of a *seta*, scale bar 5 μm . (d) High-magnification image of the regular structure in the *seta*, scale bar 2 μm (Image (a) reproduced with permission of A. R. Parker, 1995. Images (b-d) reproduced with permission of P. Vukusic and The American Physical Society © 2009 Trzeciak and Vukusic [26]). (e) The colorful spines (white arrow) of the sea mouse *Aphrodita* sp. (Image courtesy of D. Harasti). (f) Optical micrograph of one of the colorful spines. (g, h) Side view and cross-

section of the tubular structures in the spine, scale bars 2 and 1 μm (Images (f-h) reproduced with permission of A. R. Parker and The Royal Society © 2004 Parker [27]). (i) The bright green wattle (white arrow) of the Madagascan bird *Philepitta castanea* (velvet asity). (j) A transmission electron micrograph of a cross-section of the nanostructured arrays of dermal collagen fibers responsible for the wattle's green color, scale bar 500 nm. (k) The blue eye spot (white arrow) of the Madagascan bird *Terpsiphone mutata* (Madagascar paradise flycatcher). (l) A cross-section of the collagen fiber array that causes the blue color around the bird's eye, scale bar 500 nm (Images (i-l) adopted from Prum and Torres [10] with permission of R. O. Prum and T. Schulenberg (k))



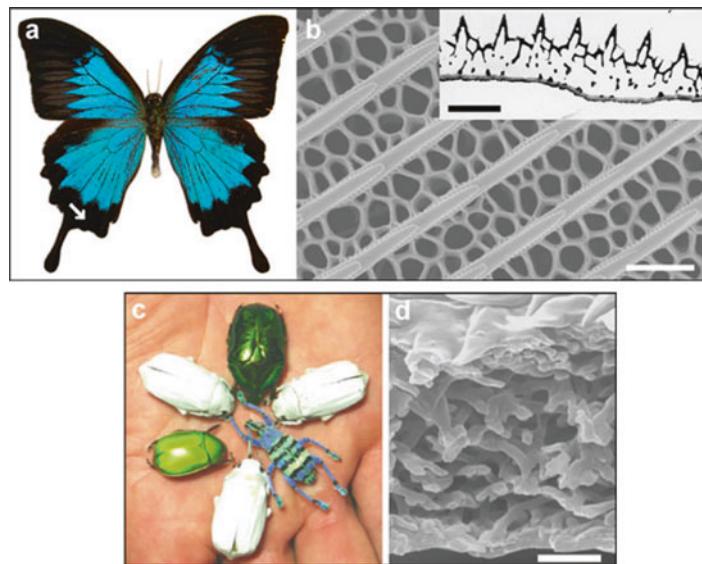
Structural Color in Animals, Fig. 4 Structural colors based on three-dimensional photonic crystals. (a) Dorsal view of the Australian weevil *Pachyrhynchus argus* showing the green metallic spots (white arrow) on its back side. (b) Scanning electron micrograph of the beetles scales, scale bar 50 μm . (c) Cross-section of a scale revealing the opaline structure responsible for the beetle's metallic coloration, scale bar 1 μm (Images (a–c) reproduced with permission of A.R. Parker and The Royal Society © 2004 Parker [27]). (d) The South-East Asian bright blue longhorn beetle *Pseudomyagrus waterhousei*. (e) Micrograph of the iridescent drop-shaped scales on the beetle's body, scale bar 50 μm . (f) Scanning electron micrograph of a cross-section through a single scale revealing the origin of the iridescent blue color, regularly sized, closely packed spherules forming a direct opal, scale bar 1 μm (Images (d–f) reproduced with permission of J. P. Vigneron and The American Physical Society © 2011 Simonis and Vigneron [29]). (g) The beetle *Pachyrrhynchus congestus pavonius*. (h) Micrograph of the highly conspicuous annular spots on

the beetle's thorax. The scales of different colors are clearly visible. (i) Scanning electron micrograph of the cross-section of a scale, revealing the photonic structure consisting of a combination of a multilayer reflector on top of a three-dimensional face-centered cubic photonic crystal, scale bar 1 μm (Images (g–i) reproduced with permission of J. P. Vigneron and The American Physical Society © 2007 Welch et al. [30]). (j) The South American butterfly *Parides sesostris*, commonly called Emerald-patched Cattleheart due to the bright green patches on its upper wing pairs (white arrow). (k) Scales in the green areas of the butterfly's wings, scale bar 100 μm . (l) SEM image of a cross-section of the three-dimensional photonic structure found within a single scale, scale bar 1 μm . (m) Transmission electron micrograph of the cross-section of a scale showing the superficial ridging and the domains of the underlying of photonic crystal, scale bar 2 μm (Images (j, l, m) reproduced with permission of P. Vukusic and Springer Science + Business Media B.V. © 2009 Vukusic [19]. Image (k) courtesy of M. Doolittle)

and color purity is ensured by absorption of undesired spectral light components in regions beneath or around the color creating photonic elements. Consequently, optimized absorption plays an important role for natural structural colors. The scales of various butterflies that display structural colors are supported on a surface containing melanin pigments. The pigments

absorb the light that passes through photonic structures of the scales, suppressing spurious reflection and consequently preventing the desaturation of the reflected color.

Very strong blackness can be achieved by an optimized interplay of pigment absorption and structure. A particularly good example is the butterfly *Papilio ulysses* [7]. The bright blue patches



Structural Color in Animals, Fig. 5 Deep blackness and brilliant whiteness. (a) The butterfly *Papilio ulysses* shows regions of deep black (white arrow) surrounding the bright blue spots on its wings. (b) The scales in the black regions carry cuticle microstructures that trap light. The inset shows a cross-section of a scale revealing the dense network of pigment-loaded cuticle, scale bars 2 μm . (c) The brilliant white beetle *Cyphochilus* spp. compared to other

beetles. (d) The intense white results from incoherent diffuse scattering from an interconnected network of filaments within the body of each beetle scale, scale bar 1 μm (Inset in (b), images (c) and (d) reproduced with permission of P. Vukusic, The Royal Society © 2004 Vukusic et al. [7] and The Optical Society of America © 2009 Hallam et al. [32])

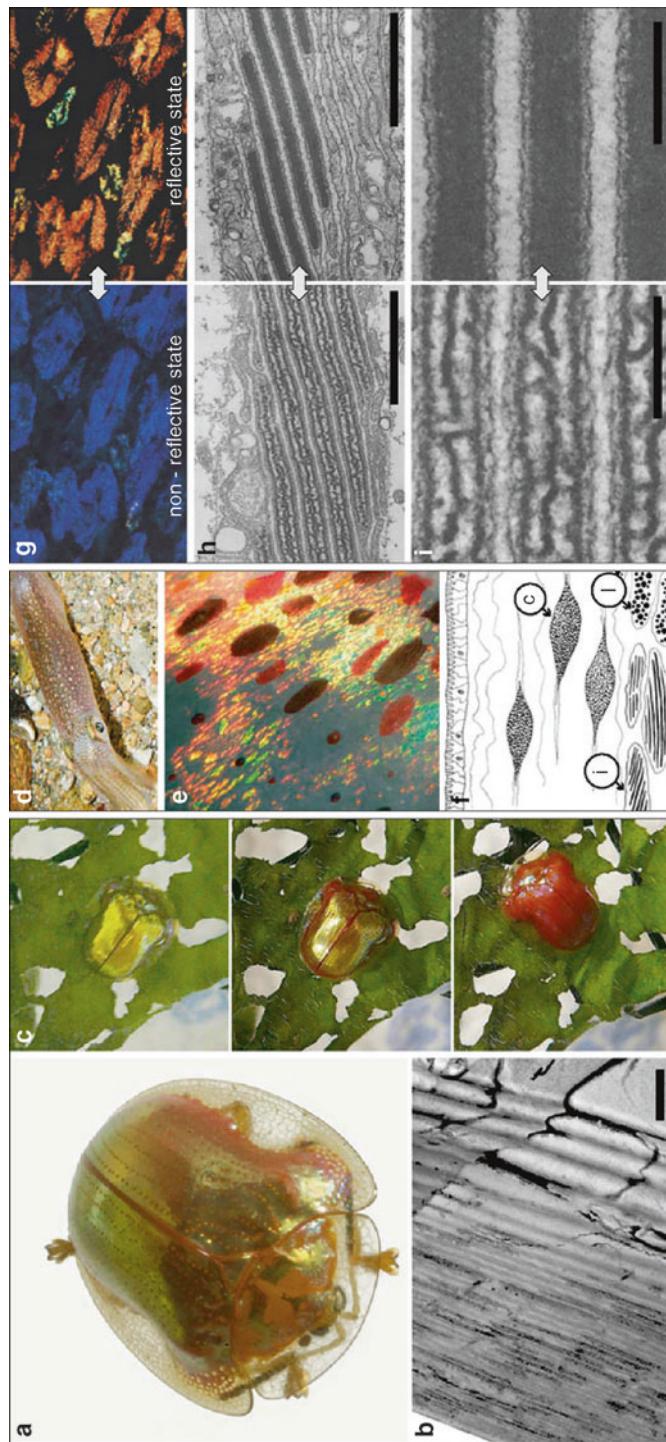
on the butterfly's wings are surrounded by very dark scales that show strong light absorption induced by melanin-containing microstructures (Fig. 5a). These pigment-filled structures act as light traps that are tailored to prevent light from escaping once it enters the scales (Fig. 5b). This maximizes the interaction with the pigment, resulting in the extremely low reflectivity of the butterfly's black scales.

Materials with a high structural irregularity, a random distribution, orientation, and size of individual transparent or highly reflective scattering elements on the scale of the wavelength of light are known to equally scatter light of any color into a broad angular range, thereby creating an intense white appearance. An extra-ordinary example of brilliant structural whiteness has been discovered in the case of the *Cyphochilus* spp. beetle [6]. The insect's intense white (Fig. 5c) is based on the incoherent scattering of light from a random network of about 250 nm thick interconnected filaments that are randomly distributed within the

ultrathin beetle scales (Fig. 5d). Despite the fact that the scales are only about 5 μm thick, an optimized void fraction, high aperiodicity, and the large refractive index contrast of about 0.56 ensure that light of all colors is very efficiently scattered in all directions.

Dynamically Variable Structural Colors in Animals

The most interesting structural color systems in nature are the ones that show aspects of intentional color tuning, like the photonic structure employed by the beetle *Charidotella egregia* (Fig. 6a). In situations of distress, for instance, caused by predator attacks [33], the beetle can change its appearance from a bright golden shine to a striking red color (Fig. 6c) within 2 min. The structure inducing this phenomenon is a multilayer buried in the exocuticle of the beetle's transparent armor (Fig. 6b). The multilayer is chirped



Structural Color in Animals, Fig. 6 Dynamically tunable structural color. (a) The golden beetle *Charidotella egregia*. (b) Transmission electron micrograph of the chirped multilayer reflector in the outer layer of the beetle's

armor, showing the density variation of irregularly distributed porous and solid areas within the stack, scale bar 1 μm . (c) When threatened the beetle can change its golden color to a striking red within less than 2 min (Images (a–c)

which results in reflection of light in a broad wavelength range, leading to the golden appearance of the beetle. Internally, the multilayer structure shows a random distribution of porous regions and channels within the layer planes and perpendicular to the layers. In the default state, these irregularly distributed voids are filled with liquid leading to a homogenous refractive index within each layer thereby suppressing scattering and allowing interference of light in the chirped multilayer, which results in the golden metallic color. Beneath the multilayer resides a layer of bright red pigment. When under a potential threat the beetle resorts to an aposematic protection mechanism. It can withdraw the liquid from the porous regions in the multilayer which strongly increases the scattering thereby rendering the layer stack translucent and revealing the bright red color of the underlying pigments, which serves as a warning signal.

The impressive camouflage and signaling capacities of cephalopods (squid, cuttlefish, and octopus) result from an intricate interplay of mainly three different functional elements within the skin of the animals [35]. Chromatophores are small pigment-filled organs that can be stretched and compressed by radially attached muscle strands. The squid *Logilo pealeii* (Fig. 6d) can vary the size of its chromatophores from 1.5 mm in the expanded state to about 0.1 mm when retracted [34]. A distinct layer of iridophores buried in the skin beneath the chromatophores provide spectrally selective reflection (Fig. 6e, f). The iridophores are colorless cells, which contain stacks of thin platelets (iridosomes) that reflect light by multilayer interference. Among the iridophores, cuttlefish and octopus employ

additional structural elements, so-called leucophores (Fig. 6f). These cells are made up of disordered spherical assemblies with particles ranging from 250 to 1,250 nm in diameter (leucosomes). They induce diffuse broadband scattering and are responsible for the white patterns on cuttlefish and octopus. While the leucophores are passive elements, cephalopods have a high physiological control over chromatophores and iridophores [35]. The chromatophores can be expanded or retracted within a fraction of a second. In the expanded state they display the pigment color (red, yellow/orange, or brown/black depending on the species) and hide the underlying iridophores and leucophores from interfering with the incident light. When the chromatophores are retracted iridophores and leucophores are revealed and determine the reflected color. Squids are able to change the iridescence of the iridophores with shifts of over 100 nm in the reflected wavelengths observed for some species. This reflectance change progresses much slower than the actuation of the chromatophores and can take several seconds to minutes. The change in iridophore reflection can result from two different processes. The platelets in the iridophores, which are made of a protein called reflectin, can change their refractive index by a change in state of the protein conformation [36] (Fig. 6g–i). Furthermore, the thickness of the plates can change to tune the reflection. In addition to the described passive optical elements, some cephalopod species make use of structurally very complex light-emitting photophores that can be highly directional in their emission. The emission of photophores is often enhanced by the incorporation of multilayer back-reflectors. More



Structural Color in Animals, Fig. 6 (continued) reproduced with permission of J.P. Vigneron and The American Physical Society © 2007 Vigneron et al. [33]. (d) The squid *Logilo pealeii*. (e) Microscope images of the squid's skin show the brown and reddish pigmented chromatophores and the underlying iridescent iridophores shimmering in colors from green to orange. (f) A sketch of the arrangement of chromatophores (c), iridophores (i), and leucophores (l) in the skin of cephalopods (Images (d–f) reproduced with permission of L. Mäthger, Springer Science + Business Media © 2007 Mäthger and Hanlon

[34] and The Royal Society © 2009 Mäthger et al. [35]). (g) Optical microscope images, and (h, i) transmission electron micrographs of active iridophores of the squid *Lolliguncula brevis* reveal the variation in reflection and the ultrastructural changes in the iridophore platelets when switching between non-reflective (*left*) and reflective (*right*) states, scale bars 1 μm (h), 250 nm (i) (Images (g–i) reproduced with permission of R. T. Hanlon and Springer Science + Business Media © 1990 Cooper et al. [36])

complex photophores contain filters and light guides that help to channel and direct the emitted light. In summary, cephalopods use an extensive repertoire of actively tunable optical elements that rely on a range of physical effects in order to manipulate incident light, including absorption, light interference, bioluminescence, and scattering, which makes them the uncontested masters of color, light manipulation, and camouflage in nature.

Conclusion

In the course of evolution, various organisms in nature have developed a huge variety of photonic systems that by interference, diffraction, coherent, and incoherent scattering cause distinct color. While the bearers of such photonic structures come from very different animal orders, common design principles can be identified across the distinct taxonomic groups. Regular periodic multilayer arrangements and two- or three-dimensional photonic crystals build the base for strong color, while disordered arrays of particles and filaments cause bright whiteness. Pigment-loaded structures with corrugations on different length scales act as efficient light traps that render surfaces deep black. Strong blackness and brilliant whiteness are frequently employed in nature to provide contrast for intriguing patterns of color on wings and bodies of insects, scales and shells of marine animals, and feathers of birds. The coordinated interplay of regularity and irregularity on different length scales plays an important role in the function of many natural photonic systems [4, 9]. Well-defined, structural regularity and periodicity on the submicron scale ensures the reflection of strong bright colors, while irregularity on the scale of several microns inducing random scattering often mediates color stability and conspicuousness in a wide angular range.

By looking at organisms in nature new insight and knowledge can be gained for the design of materials that show specific and efficient interaction with light. This fact is widely acknowledged in the scientific community [15, 27]. Researchers show increased interest in the development of bio-inspired photonic systems. The current

techniques and tools used in the industry, research, and everyday life for light harvesting, optical signaling, data transfer, and processing might soon benefit from a better understanding of the composition and functioning of biological photonic structures.

Cross-References

- [Biomimetics of Optical Nanostructures](#)
- [Moth-Eye Antireflective Structures](#)
- [Nanostructures for Coloration \(Organisms Other than Animals\)](#)
- [Nanostructures for Photonics](#)

References

1. Kinoshita, S., Yoshioka, S., Miyazaki, J.: Physics of structural colors. *Rep. Prog. Phys.* **71**, 076401 (2008)
2. Land, M.F.: The physics and biology of animal reflectors. *Prog. Biophys. Mol. Biol.* **24**, 75–106 (1972)
3. Parker, A.R., Martini, N.: Structural colour in animals – simple to complex optics. *Opt. Laser Technol.* **38**, 315–322 (2006)
4. Kinoshita, S., Yoshioka, S.: Structural colors in nature: the role of regularity and irregularity in the structure. *Chemphyschem* **6**, 1442–1459 (2005)
5. Vukusic, P., Sambles, J.R.: Photonic structures in biology. *Nature* **424**, 852–855 (2003)
6. Vukusic, P., Hallam, B., Noyes, J.: Brilliant whiteness in ultrathin beetle scales. *Science* **315**, 348 (2007)
7. Vukusic, P., Sambles, J.R., Lawrence, C.R.: Structurally assisted blackness in butterfly scales. *Proc. R. Soc. B* **271**, S237–S239 (2004)
8. Vukusic, P., Stavenga, D.G.: Physical methods for investigating structural colours in biological systems. *J. R. Soc. Interface* **6**, S133–S148 (2009)
9. Kinoshita, S.: Structural Colors in the Realm of Nature. World Scientific, Singapore (2008)
10. Prum, R.O., Torres, R.: Structural colouration of avian skin: convergent evolution of coherently scattering dermal collagen arrays. *J. Exp. Biol.* **206**, 2409–2429 (2003)
11. Prum, R.O., Torres, R.H., Williamson, S., Dyck, J.: Coherent light scattering by blue feather barbs. *Nature* **396**, 28–29 (1998)
12. Parker, A.R.: 515 million years of structural colour. *J. Opt. A Pure Appl. Opt.* **2**, R15–R28 (2000)
13. Parker, A.R.: A geological history of reflecting optics. *J. R. Soc. Interface* **2**, 1–17 (2005)
14. Prum, R.O., Torres, R.H.: Structural colouration of mammalian skin: convergent evolution of coherently scattering dermal collagen arrays. *J. Exp. Biol.* **207**, 2157–2172 (2004)

15. Biró, L.P., Vigneron, J.P.: Photonic nanoarchitectures in butterflies and beetles: valuable sources for bioinspiration. *Laser Photonics Rev.* **5**, 27–51 (2011)
16. Ghiradella, H.: Shining armor: structural colors in insects. *Opt. Photonics News* **10**, 46–48 (1999)
17. Seago, A.E., Brady, P., Vigneron, J.P., Schultz, T.D.: Gold bugs and beyond: a review of iridescence and structural colour mechanisms in beetles (*Coleoptera*). *J. R. Soc. Interface* **6**, S165–S184 (2009)
18. Srinivasarao, M.: Nano-optics in the biological world: beetles, butterflies, birds, and moths. *Chem. Rev.* **99**, 1935–1962 (1999)
19. Vukusic, P.: Advanced photonic systems on the wing-scales of *Lepidoptera*. In: Gorb, S.N. (ed.) *Functional Surfaces in Biology: Little Structures with Big Effects*. Springer, Dordrecht (2009)
20. Doucet, S., Meadows, M.: Iridescence: a functional perspective. *J. R. Soc. Interface* **6**, S115–S132 (2009)
21. Yoshioka, S., Kinoshita, S.: Polarization-sensitive color mixing in the wing of the Madagascan sunset moth. *Opt. Express* **15**, 2691–2701 (2007)
22. Vukusic, P., Sambles, J.R., Lawrence, C.R.: Structural colour: colour mixing in wing scales of a butterfly. *Nature* **404**, 457 (2000)
23. Parker, A.R., Hegedus, Z.: Diffractive optics in spiders. *J. Opt. A Pure Appl. Opt.* **5**, S111–S116 (2003)
24. Vigneron, J.P., et al.: Reverse color sequence in the diffraction of white light by the wing of the male butterfly *Pierella luna* (Nymphalidae: Satyrinae). *Phys. Rev. E* **82**, 021903 (2010)
25. Parker, A.R., McPhedran, R.C., McKenzie, D.R., Botten, L.C., Nicorovici, N.P.: Aphrodites iridescence. *Nature* **409**, 36–37 (2001)
26. Trzeciaik, T.M., Vukusic, P.: Photonic crystal fiber in the polychaete worm *Pherusa* sp. *Phys. Rev. E* **80**, 061908 (2009)
27. Parker, A.R.: A vision for natural photonics. *Philos. Trans. R. Soc. A* **362**, 2709–2720 (2004)
28. Parker, A.R., Welch, V.L., Driver, D., Martini, N.: Structural colour: opal analogue discovered in a weevil. *Nature* **426**, 786–787 (2003)
29. Simonis, P., Vigneron, J.P.: Structural color produced by a three-dimensional photonic poly-crystal in the scales of a longhorn beetle: *Pseudomyagrus waterhousei* (Coleoptera: Cerambicidae). *Phys. Rev. E* **83**, 011908 (2011)
30. Welch, V., Lousse, V., Deparis, O., Parker, A.R., Vigneron, J.P.: Orange reflection from a three-dimensional photonic crystal in the scales of the weevil *Pachyrrhynchus congestus pavonius* (Curculionidae). *Phys. Rev. E* **75**, 041919 (2007)
31. Saranathana, V., et al.: Structure, function, and self-assembly of single network gyroid ($I_4,32$) photonic crystals in butterfly wing scales. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 11676–11681 (2010)
32. Hallam, B.T., Hiorns, A.G., Vukusic, P.: Developing optical efficiency through optimized coating structure: biomimetic inspiration from white beetles. *Appl. Opt.* **48**, 3243–3249 (2009)
33. Vigneron, J.P., et al.: Switchable reflector in the Panamanian tortoise beetle *Charidotella egregia* (Chrysomelidae: Cassidinae). *Phys. Rev E* **76**, 031907 (2007)
34. Mäthger, L.M., Hanlon, R.T.: Malleable skin coloration in cephalopods: selective reflectance, transmission and absorbance of light by chromatophores and iridophores. *Cell Tissue Res.* **329**, 179–186 (2007)
35. Mäthger, L.M., Denton, E.J., Marshall, N.J., Hanlon, R.T.: Mechanisms and behavioural functions of structural coloration in cephalopods. *J. R. Soc. Interface* **6**, S149–S163 (2009)
36. Cooper, K.M., Hanlon, R.T., Budelmann, B.U.: Physiological color change in squid iridophores. II. Ultrastructural mechanisms in *Lolliguncula brevis*. *Cell Tissue Res.* **259**, 15–24 (1990)

Structural Colors

► [Nanostructures for Coloration \(Organisms Other Than Animals\)](#)

Structural DNA Nanotechnology

► [DNA Origami as Programmable Nanofabrication Tools](#)

Structural Fluctuations

► [DNA from First Principles](#)

Structure and Stability of Protein Materials

Szu-Wen Wang
Chemical Engineering and Materials Science,
The Henry Samueli School of Engineering,
University of California, Irvine, CA, USA

Synonyms

[Nanostructure](#); [Thermostability](#)

Definition

Protein-based materials are polymeric biomaterials comprising amino acid subunits that are connected together by peptide bonds. These materials are usually biomimetic, self-assemble into higher-order nanometer-scale architectures, and can interact with biological entities. Determination of their structure and stability is an important component of assessing their utility and function.

Protein-Based Nanomaterials

The control of architecture at the nanoscale is a challenge in which nature has been highly successful. Since genetic manipulation enables the definition of every monomer in a polymeric protein structure, giving far greater control than conventional chemical synthesis, one approach in material synthesis is the use of protein engineering to create biologically inspired materials [1]. By combining natural scaffolds, structural elements, and biologically reactive sites, materials with novel architectures and properties can be obtained. These materials are fabricated in microorganisms as recombinant proteins comprising amino acid units. Furthermore, proteins often self-assemble into complex, higher-order nanostructures, which include fibrous and spherical architectures [2, 3]. Protein-based materials have been evaluated for a wide variety of applications, such as tissue engineering scaffolds, templates for nanomaterial synthesis, and drug delivery carriers. The evaluation of structure and stability in these protein-based nanomaterials is primarily established using techniques developed for general proteins.

Structure

The structures of proteins span several length scales, and different methods are used depending on the type of structure being evaluated [4]. The *primary structure* of proteins describes the sequence of individual amino acids, or monomer

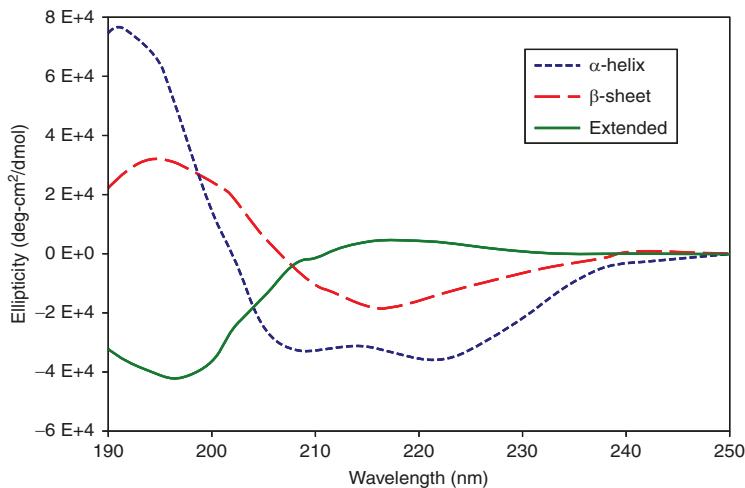
residues, covalently coupled by peptide bonds to form a polypeptide chain. There are 20 known natural amino acids, each differing in their side chain identity, thereby imparting different physicochemical characteristics at each location along the polymer. One can also incorporate “unnatural” amino acids, which expands the functionalization of the protein-based material with useful chemical components that are not native [5, 6]. When the protein materials have been created in a recombinant system (such as in microorganisms), the DNA encodes for the protein sequence. Primary structure can also be determined by N-terminal peptide sequencing or proteolytic mass spectrometry.

The local primary structure affects the *secondary structure*, which are regions of local folding. Secondary structure of protein materials is commonly determined by circular dichroism (CD) [7]. Asymmetric molecules, such as polypeptides, will absorb left and right circularly polarized light to different degrees. Circular dichroism is this absorbance difference, and different types of secondary folding (such as α -helices, β -sheets, or extended helical conformations) have characteristic CD spectra (Fig. 1, [8]). The degree of folding can be used to monitor conformational changes of protein materials.

Tertiary structure is the three-dimensional conformation of the folded protein. The *quaternary structure* is the assembly of the individual polypeptide or protein subunits. Global size or assembly properties, such as hydrodynamic diameters or molecular weights of the protein complexes, can be evaluated by dynamic light scattering or ultracentrifugation. While these techniques can give useful information, they do not elucidate structural details. The most precise structural information can be obtained by techniques such as X-ray diffraction or nuclear magnetic resonance spectroscopy. However, these methods require highly pure protein, a crystalline form of the protein, and/or extensive data analysis, and these conditions are not usually available or feasible for protein-based materials used in nanotechnology applications. Therefore, alternative methods providing intermediate amounts of information are used to probe tertiary and quaternary structure.

Structure and Stability of Protein Materials,

Fig. 1 Circular dichroism spectra of polypeptides with representative α -helix (dotted line), β -sheet (dashed line), and extended helical (solid line) conformations (Data replotted from Ref. [8]). Reprinted with permission, copyright 1969, American Chemical Society)



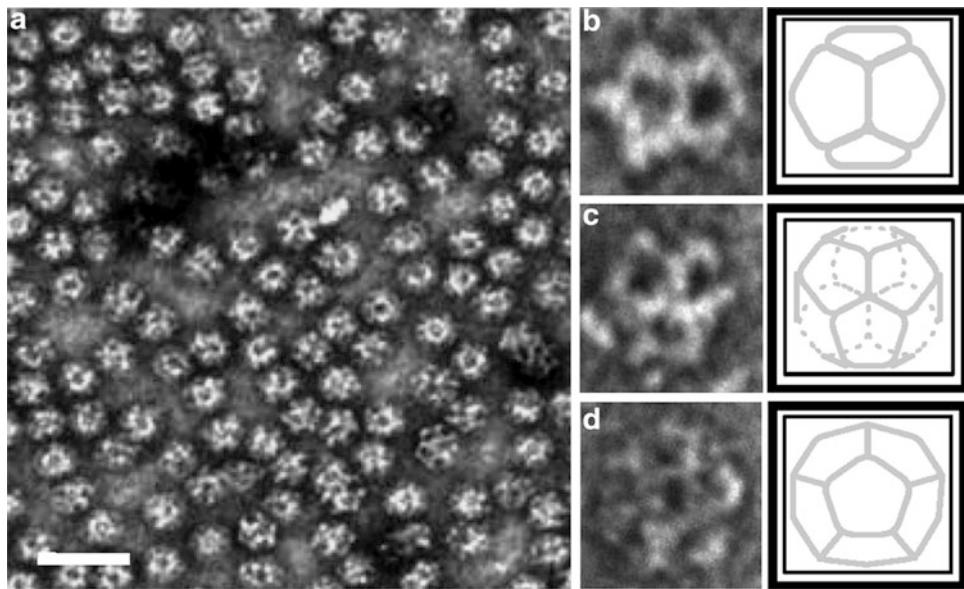
Microscopy strategies enable visualization of protein nanostructures ranging from nanometer-to-micron length scales. In transmission and scanning transmission electron microscopy (TEM and STEM, respectively), an electron beam is transmitted through a sample, and information regarding the scattered electrons is obtained [9]. Proteins are first deposited in a thin layer onto carbon-coated electron microscope grids. These grids are then frozen at liquid nitrogen temperatures in cryogenic preparation, or proteins can also be negatively stained with a heavy ion salt (e.g., uranyl acetate) to provide contrast. Due to the insulating nature of proteins, the resolution using TEM or STEM is low relative to conducting samples and is typically on the order of 5–10 nm (Fig. 2, [10]).

Alternatively, atomic force microscopy (AFM) can yield lateral resolutions of less than 1 nm for protein structures (Fig. 3, [9]). Proteins are adsorbed onto an atomically flat substrate (typically mica or graphite) and are imaged dried or hydrated in buffer. As the AFM probe is scanned over the surface of a sample, deflections (due to interactions between the probe tip and sample) are measured. Wet samples reflect native protein conditions, and therefore AFM potentially can yield more accurate structural information relative to electron microscopy. One challenge of AFM relative to electron microscopy, however, is the time-intensive nature of imaging.

Stability and Mechanical Properties

The stability of protein materials can be evaluated as the structure and assembly are probed while stressing conditions (such as temperature, pH, denaturants, or ionic strength) are changed. For example, the thermostability of protein materials can be evaluated by CD measurement at a characteristic wavelength as the temperature of the sample is increased [11]. Using this data, thermodynamic and thermostability values for protein unfolding (e.g., enthalpy, entropy, free energy, midpoint of unfolding temperature, and onset of unfolding temperatures) can be calculated. Differential scanning calorimetry (DSC), which measures the heat capacity of a protein solution as a function of temperature, can also determine unfolding and thermodynamic parameters of protein-based materials [12]. DSC can also be used to gain insight into kinetic stability and conditions in which protein materials exist at a local energetic minimum with a nonideal structure [13]. Although more direct thermodynamic information can be obtained from DSC, advantages of CD over DSC include the significant lower amounts of protein required for analysis and greater accessibility to spectropolarimeters.

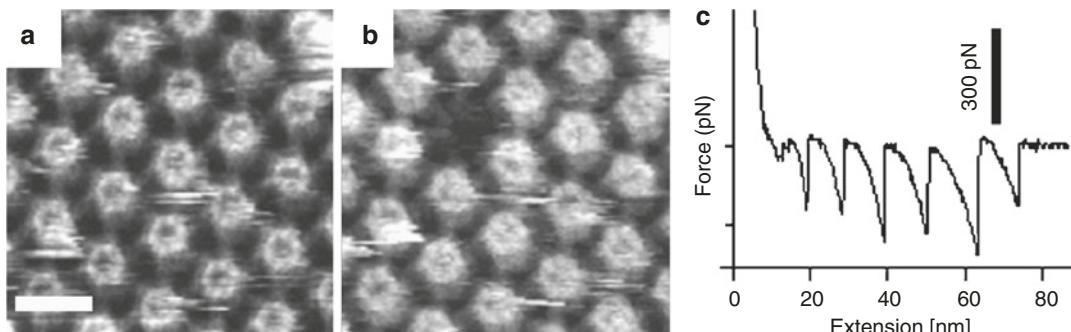
Applying force to single protein molecules gives insight into how the materials may be assembled and the stability of these interactions. Using single-molecule force spectroscopy, AFM can be used to pull out proteins within structured



Structure and Stability of Protein Materials,

Fig. 2 Transmission electron microscopy image of dodecahedral protein complex. (a) Negatively stained protein assembly of the E2 subunit from pyruvate

dehydrogenase. Scale bar is 50 nm. Projection views of (b) twofold, (c) threefold, and (d) fivefold axes of symmetry verify icosahedral assembly (Figure from Ref. [10]. Reprinted with permission, copyright 2008, Wiley)



Structure and Stability of Protein Materials,

Fig. 3 Atomic force microscopy (AFM) analyses of protein structures. (a) Polymers of recombinant human collagen, with each strand corresponding to a trimeric assembly. Image size is $1 \times 1 \mu\text{m}$ (Image provided by Richard Que, University of California, Irvine). (b) Extracellular

hexagonally packed intermediate (HPI) monolayer. Scale bar is 20 nm. (c) Force-distance curve of a HPI monolayer exhibits six peaks, corresponding to successive unfolding of the linked HPI proteins. (Panels b and c are from Ref. [9]. Reprinted with permission, copyright 2011, Elsevier)

assemblies, and force-distance curves yield information about the forces required to disrupt folded proteins and protein complexes [9, 14]. Subsequent imaging by AFM can show the resulting effects on the complex (Fig. 3, [9]).

Protein-based materials often demonstrate both viscous and elastic properties, particularly if they

exist at high concentrations or as hydrogels. The storage modulus (G') describes the elastic, solid-like properties, and the loss modulus (G'') measures its viscous, liquid-like character. These dynamic moduli can be determined experimentally with a rheometer or with particle-tracking microrheology [15]. The latter technique

measures the displacement of microparticles mixed with protein solution over time and is particularly useful when only small amounts of protein are available, which is often the case in early development of protein materials. Static mechanical properties, such as tensile stress and tensile strain, of bulk protein-based hydrogels can be measured using a conventional mechanical tester.

Cross-References

- [AFM in Liquids](#)
- [AFM, Tapping Mode](#)
- [Atomic Force Microscopy](#)
- [Bioinspired Synthesis of Nanomaterials](#)
- [Biomimetic Synthesis of Nanomaterials](#)
- [Biomimetics](#)
- [Biomimetics of Marine Adhesives](#)
- [Mechanical Properties of Hierarchical Protein Materials](#)
- [Spider Silk](#)
- [Transmission Electron Microscopy](#)

References

1. DiMarco, R.L., Heilshorn, S.C.: Multifunctional materials through modular protein engineering. *Adv. Mater.* **24**, 3923–3940 (2012)
2. Kluge, J.A., Rabotyagova, U., Leisk, G.G., Kaplan, D. L.: Spider silks and their applications. *Trends Biotechnol.* **26**, 244–251 (2008)
3. Molino, N.M., Wang, S.W.: Caged protein nanoparticles for drug delivery. *Curr. Opin. Biotech.* **28**, 75–82 (2014)
4. Creighton, T.E.: Proteins: Structures and Molecular Properties, 2nd edn. Freeman, New York (1993)
5. Xie, J.M., Schultz, P.G.: Innovation: a chemical toolkit for proteins – an expanded genetic code. *Nat. Rev. Mol. Cell Biol.* **7**, 775–782 (2006)
6. Connor, R.E., Tirrell, D.A.: Non-canonical amino acids in protein polymer design. *Polym. Rev.* **47**, 9–28 (2007)
7. Greenfield, N.J.: Using circular dichroism spectra to estimate protein secondary structure. *Nat. Protoc.* **1**, 2876–2890 (2006)
8. Greenfield, N.J., Fasman, G.D.: Computed circular dichroism spectra for evaluation of protein conformation. *Biochemistry* **8**, 4108–4116 (1969)
9. Muller, S.A., Muller, D.J., Engel, A.: Assessing the structure and function of single biomolecules with scanning transmission electron and atomic force microscopes. *Micron* **42**, 186–195 (2011)
10. Dalmau, M., Lim, S., Chen, H.C., Ruiz, C., Wang, S. W.: Thermostability and molecular encapsulation within an engineered caged protein scaffold. *Biotechnol. Bioeng.* **101**, 654–664 (2008)
11. Greenfield, N.J.: Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. *Nat. Protoc.* **1**, 2527–2535 (2006)
12. Ladbury, J.E., Doyle, M.L. (eds.): Biocalorimetry 2: Applications of Calorimetry in the Biological Sciences. Wiley, Hoboken (2004)
13. Sanchez-Ruiz, J.M.: Protein kinetic stability. *Biophys. Chem.* **148**, 1–15 (2010)
14. Muller, D.J., Dufrene, Y.F.: Atomic force microscopy as a multifunctional molecular toolbox in nanobiotechnology. *Nat. Nanotechnol.* **3**, 261–269 (2008)
15. Wirtz, D.: Particle-tracking microrheology of living cells: principles and applications. *Annu. Rev. Biophys.* **38**, 301–326 (2009)

Structure of Nanoparticles

- [High Energy Synchrotron Radiation and Its Impact on Characterizing Nanoparticles](#)

SU-8 Photoresist

Frederik Ceyssens and Robert Puers
Department ESAT-MICAS, KULeuven, Leuven, Belgium

Synonyms

Gamma-butyrolactone (GBL); Hardbake (HB); Lithographie, galvanoformung, abformung (LIGA); *N*-Methyl-2-pyrrolidone (NMP); Polyethylene carbonate (PEC); Polymethyl methacrylate (PMMA); Polypropylene carbonate (PPC); Postexposure bake (PEB); Propylene glycol methyl ether acetate (PGMEA)

Definition

SU-8 is a high aspect ratio epoxy-based negative photoresist commonly used as structural material in lithographic fabrication.

Introduction

SU-8 was developed by IBM as a thick negative photoresist targeted to the fabrication of molds for electroplating. The epoxy-based negative photoresist has some remarkable properties. First of all, it has a wide range of coating thicknesses: layers from several hundreds of nanometers up to several hundreds of microns can be deposited by a single standard spin coating step, using the appropriate dilution of the SU-8 resin. Even thicker layers can be deposited and photopatterned successfully as will be discussed later. Second, structures featuring almost straight sidewalls can be created in the SU-8 layer by simple UV exposure through a contact mask. Aspect ratios of about 15 are routinely obtained. Using optimized processes in layers over a few hundreds of microns thick, aspect ratios of over 100 are obtainable [1].

The fabrication of such polymer structures was up to that time only possible by the expensive LIGA technique which requires synchrotron radiation. Though LIGA is still superior in terms of achievable aspect ratio and wall straightness, SU-8 can be humoristically referred to as a “poor man’s LIGA.” This must not be taken pejoratively per se as “cheap” can be read as “opening new possibilities” just as well.

SU-8 has high chemical resistance and because of the high chemical functionality of the monomers its mechanical properties are good enough to render it useful as structural material in a large number of cases, as further discussed in the material properties section.

SU-8 is an insulator, but covered later it can be made conductive, for example, by metal deposition. Also, it can of course be used as an electroplating mold to create thick, high aspect ratio metal structures.

It is transparent, enabling easy inspection of the underlying structures and certain optical

readout principles, for example, based on fluorescence. Furthermore, the processing of SU-8 does not require a large investment in equipment or consumables.

This all leads to the fact that SU-8 has undoubtedly become one of the most widely used structural polymer materials in microsystems. It is not only frequently used in work of the traditional micromechanics community but has also spread into groups working on labs-on-chip, integrated optics, packaging, and others. This is illustrated by the large number of research papers mentioning SU-8 in the title that have appeared over the last decade (Fig. 1).

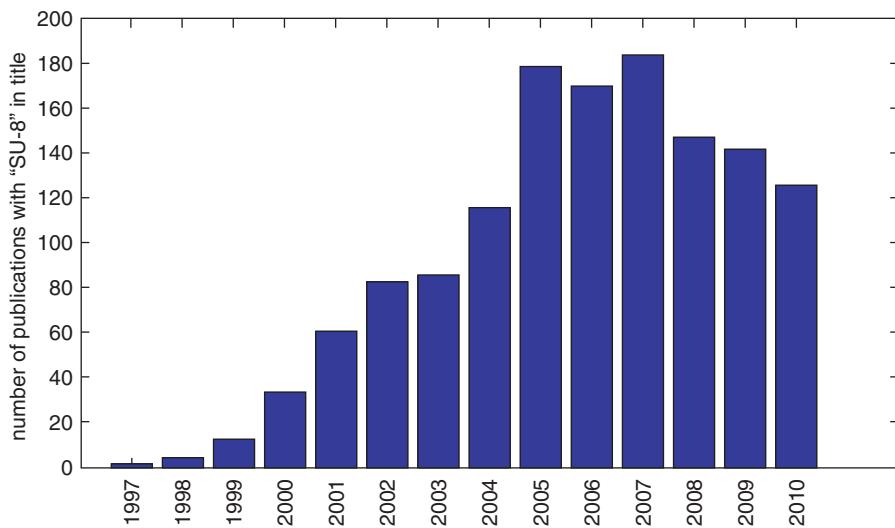
Commercially available SU-8 resist has a solid fraction consisting of SU-8 monomer (Fig. 2), which contains eight epoxy functional groups, and a mixture of triaryl sulfonium hexafluoroantimonate salts as photoinitiator (4.8 wt%).

The solvent added hereto is either gamma-butyrolactone (GBL) or cyclopentanone. Both types are available from Microchem in a wide range of dilutions under the names “SU-8” or “SU-8 2000,” respectively. Another supplier is Gersteltec.

Material Properties

The mechanical and electrical properties of SU-8 are not unlike those of a typical epoxy. Some material data are summarized in Table 1. It can be seen that creep is rather low, which is a good indication of the suitability of the material for mechanical applications.

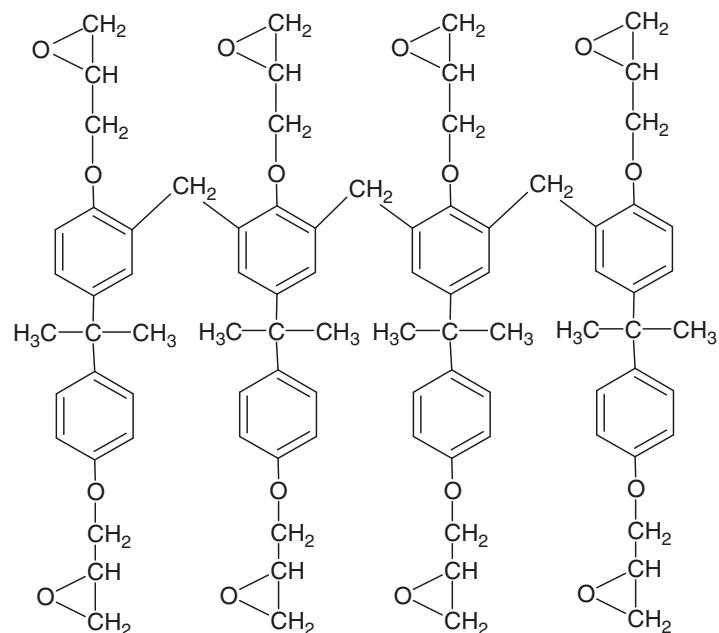
For microfluidic applications, the contact angle of the material is often important. Untreated SU-8 is rather hydrophobic, displaying a contact angle of 75° with water. The material can be made more hydrophilic by several means such as a short exposure to oxygen plasma. A treatment at 600 W for 8 s at a pressure of 27 Pa reduces the contact angle to 25.6°. After 4 min at 400 W, the contact angle is down to 3.2° but goes up again to 25° over a period of 40 days [4]. Another way is the addition of surfactants to the SU-8 itself. Bohl et al. [5] used 10 wt% trisiloxane alkoxylate to obtain a contact angle of 28°. Combined with



SU-8 Photoresist, Fig. 1 Number of research papers with “SU-8” in their title per calendar year according to Google Scholar on February 15, 2011. As SU-8 is

becoming a mature and commonly used technology, it being mentioned in the title of research papers is expected to become less frequent again

**SU-8 Photoresist,
Fig. 2** SU-8 monomer



plasma activation this becomes a mere 10°. Mostly, adhesion is sufficient (Table 2), except when using glass substrates. An adhesion promoter based on the deposition of a titanium oxide monolayer (such as AP-300) can then be used.

Processing of Thick High Aspect Ratio Structures

The procedure to deposit a single layer of SU-8 comprises seven major steps. In each of these steps, several parameters can be selected, each

SU-8 Photoresist, Table 1 Material properties of SU-8 [2]. Creep and fracture strength data from [3] who are using the following modified Voight-Kelvin model to model creep in SU-8: $\frac{\dot{\varepsilon}(t)}{\sigma(\text{applied})} = D_0 + (D_e - D_0) \left(1 - e^{-(t/\tau)^m}\right)$. With ε as the mechanical strain and t time. The mechanical stress σ_{applied} used in these experiments was 13.2 MPa

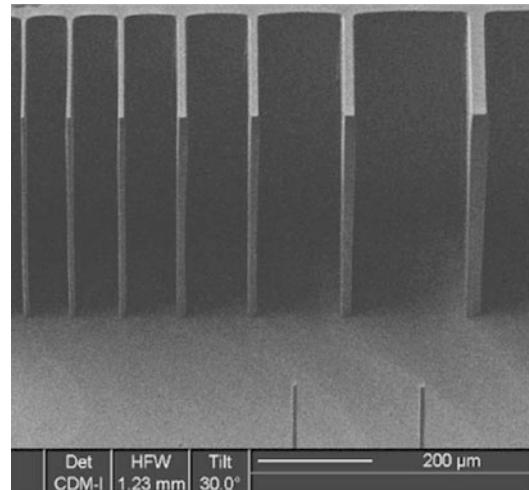
Property	Value	Remarks
Young's modulus (GPa)	2.1 ± 0.2	HB @ 150 °C for 1 h
Maximum strain to break (%)	1.2 ± 0.5	HB @ 150 °C for 1 h
Fracture strength (MPa)	73.3 ± 9.2	HB @ 180 °C for 30'
Transient creep compliance D_e (GPa $^{-1}$)	$0.310 \pm 2.5\%$	HB @ 180 °C for 30'
Initial creep compliance D_0 (GPa $^{-1}$)	$0.251 \pm 13.8\%$	HB @ 180 °C for 30'
Creep time constant τ (s)	$1.84 \times 10^{10} \pm 3.5\%$	HB @ 180 °C for 30'
Creep coupling parameter m	$0.208 \pm 5.3\%$	HB @ 180 °C for 30'
CTE (ppm/K)	52 std. 5.1	After PEB @ 95 °C
Thermal conductivity (W/m K)	0.3	
Breakdown strength (V/m)	200	Tested on a 3 μm thick layer
Refraction index	1.658	$\lambda = 365$ nm
	1.638	$\lambda = 405$ nm
Contact angle with water	75°	Untreated

having an effect on one or more properties of the final structure. Especially to attain thick, high aspect ratio features, such as shown in Fig. 3, careful processing is mandatory.

The main process sequence and its most significant parameters will be discussed below. Most important in process design is attaining a high ratio between the mechanical strength and the built-in mechanical stress in order to avoid

SU-8 Photoresist, Table 2 Shear adhesion test for SU-8 2000 [6]. Test conditions: RCA clean for glass, plasma clean for other surfaces. Test tool: Dage 4,000 working on $100 \times 100 \times 50$ μm SU-8 posts. SU-8 was hardbaked at 150 °C

Substrate	Adhesion (MPa)
Si	53
SiN	43
Ni	45
Au	29
Al/Cu (99-1)	23
Cu	38
Cu with AP-300	56
Glass	Poor
Glass with HDMS	Poor
Glass with AP-300	92
Quartz	61



SU-8 Photoresist, Fig. 3 SU-8 test structures, 700 μm high, down to 10 μm wide

deformation of the fabricated structures. This is done by controlling process parameters such as the solvent concentration during the softbake step and a few additions to the standard process such as the incorporation of relaxation steps and ramped heating [2].

Substrate Preparation

Standard cleaning procedures and a dehydration bake are recommended. On silicon, a HF dip improves adhesion and uniformity [7].

The substrate type has a large influence on uniformity and adhesion as well (Table 2).

Layer Deposition

The deposition of a uniform film of non-cross-linked SU-8 on the surface is of course of great importance for the successful outcome of the entire process. Not only is the uniformity of the height of the structures produced determined by this step, the lateral uniformity can be affected as well: an uneven distance between resist surface and mask changes the diffraction conditions during contact exposure, the most common method of UV exposing SU-8.

Spin coating is the most straightforward deposition technique for SU-8. During spinning, irregularities are introduced. The largest and always appearing irregularity is the so-called edge bead. This is a ring near the edge of the wafer where the resist is thicker. The edge bead can be negligible for thinner ($<10\text{ }\mu\text{m}$) SU-8 layers but can be over a centimeter for thick ($>100\text{ }\mu\text{m}$) layers. In the latter case, the edge bead is typically over $50\text{ }\mu\text{m}$ higher than the inner resist layer. Smaller irregularities disappear when the SU-8 layer is given enough time to reflow, preferably in a solvent-saturated atmosphere. The edge bead can be removed mechanically or by a directed spray of a solvent, a feature common in modern coating equipment.

Casting is another method commonly used to deposit SU-8 layers. During casting, a known weight or volume of low viscosity SU-8 is poured on a wafer resting on a leveled hot plate. Reflow and resist spreading in a solvent-rich environment (e.g., by putting a Petri dish over the wafer) is typically necessary. This way, for layers over a few hundreds of microns thick, superior thickness control and uniformity can be achieved. Also, no edge bead formation occurs.

In order to determine the final layer thickness, the shrinkage of the material due to solvent evaporation in the softbake step following the casting step has to be accounted for. A cast layer of SU-8 2007 was determined experimentally to shrink 44 % after a softbake of 6 h at $65\text{ }^\circ\text{C}$ followed by 7 h at $95\text{ }^\circ\text{C}$. For SU-8 2010, this is about 32 % and for SU-8 2002, 61 % [2].

Exceptionally thick layers ($>1\text{ mm}$) can be deposited by casting too. However, in such layers a significant solvent concentration gradient was observed to remain after the subsequent softbake step. This tends to weaken fine structures. It was observed that depositing and softbaking several layers on top of each other does not alleviate this significantly, as solvent from an upper layer tends to diffuse into the dry baked lower layers. A better solution is to put a known mass of predried SU-8 flakes on the wafer surface, and melt it into a smooth layer [8].

Softbake

The purpose of the softbake is to lower the solvent concentration, prevent adhesion to a contact mask, and improve the lithographic performance, that is, the ability to produce high aspect ratio structures with straight sidewalls, by reducing the diffusion of the photoinitiator. The softbake temperature must be below $120\text{ }^\circ\text{C}$, which is the onset of thermally induced cross-linking.

A too high solvent content causes the photoacid generated during exposure to diffuse easily to unexposed regions, creating bulges or protrusions in the structure. A too low solvent concentration can cause cracks and peeling off of structures.

It is, therefore, advisable to tune softbake time and temperature to reach a certain solvent concentration. In thick layers, the solvent concentration can be easily monitored by weighing the wafer on a balance with milligram precision. An advised solvent concentration is 5–7 %. An even lower solvent content decreases cross-linking [9] and increases stress deformations in high aspect ratio features [2]. After softbaking, a relaxation period of several hours is advised.

Exposure

Typically, SU-8 is exposed using contact exposure on a mask aligner equipped with a standard mercury arc UV lamp. During exposure, the photoinitiator's triarylsulfonium hexafluoroantimonate salts are split, releasing Lewis acid [7]. The acid generated serves as a catalyst, allowing a two-step reaction to occur that links the epoxy groups of different SU-8 monomers

together. In practice, the latter two-step reaction occurs only very slowly at room temperature. To speed up the reaction rate, the SU-8 is heated up after exposure, during the postexposure bake (PEB) step.

For any practical application, both the exposure dose and the wavelength of the light used are important. From the critical dose up, parts that are unsolvable in developer will be formed. The critical dose for i-line (365 nm) exposure was determined to be $30 \pm 0.5 \text{ mJ/cm}^2$ [1], for a PEB of 6' at 65 °C and 3' at 95 °C. This dose was observed to be by far insufficient for a good lithographic result: The *adhesion* of the structures formed to the substrate becomes only sufficiently strong to survive development at a much higher exposure dose. A second problem of a low dose is the permanent *deformation* of vulnerable and weakly cross-linked structures caused by internal stress during the development step. On the other hand, it is not possible to increase the exposure dose indefinitely as two problems arise with overexposure: the broadening of the features compared to their designed sizes and the appearance of effects caused by stress in larger parts. Careful choice of process parameters is therefore required. A typical i-line exposure dose is 200 mJ/cm².

Furthermore, the choice of exposure wavelength is important. Some UV wavelengths produced by a standard pressurized mercury arc lamp are absorbed too quickly and are not suitable to expose a relatively thick SU-8 layer evenly (Table 3: absorbance). The transmission of light down to a certain depth is illustrated in Fig. 2. This shows that for even exposure of a layer thicker than a few microns, the 313-nm line should always be filtered out. For layers thicker than a

few hundred microns, filtering out all light below 400 nm is recommended. As the photoinitiator is less sensitive for 405-nm light, the recommended exposure dose is higher. A dose of 10 J/cm² suffices for most applications.

When exposing layers of irregular thickness, the use of an index matching fluid such as glycerin to fill up the air gap between mask and wafer can be beneficial to reduce diffraction effects and increase resolution (Fig. 4) [24].

Postexposure Bake (PEB)

The baking step following the exposure of the resist causes a cross-linking reaction in the exposed parts, rendering those parts insoluble during the subsequent development step. Meanwhile, a few side effects occur. These effects are shrinking of the resist by the cross-linking reaction, the thermal expansion of the resist during baking, and the (dissimilar) thermal expansion of the substrate during baking. The situation is clearly different for the in-plane and the out-of-plane direction.

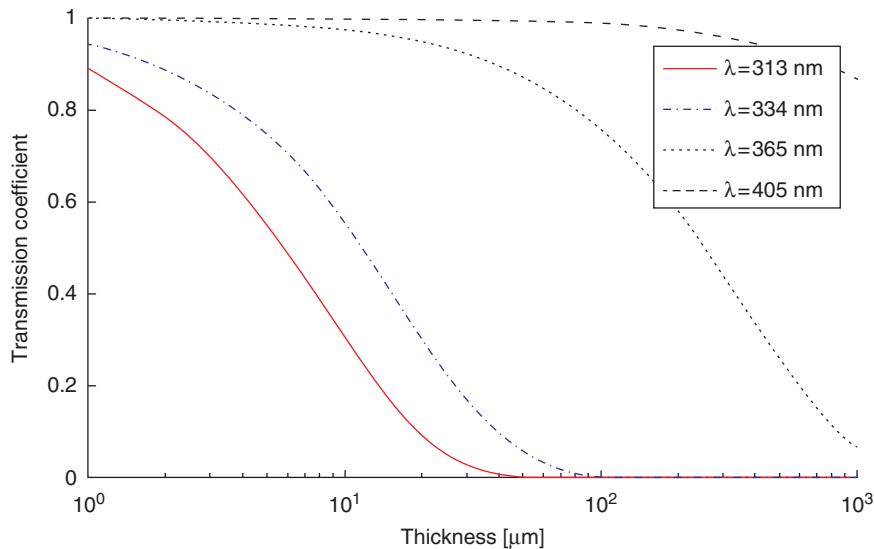
In the out-of-plane direction, the SU-8 layer is unconstrained and shrinking can occur freely. A shrinkage of approximately 4.5 % (std 0.25 %) was measured over a PEB of 1 h at 65 °C [2]. In-plane, although some relaxation occurs, an internal stress is built up (~4 MPa for a PEB temperature of 95 °C in a standard process [2]).

The combination of exposure energy and PEB time and temperature must cater for enough cross-linking for the structures to resist delamination and deformation during the subsequent development step. Typically, postexposure baking raises the glass transition temperature, that is, the temperature above which plastic deformation is possible, to a value slightly larger than the PEB temperature. After a PEB of 20 min, the glass transition temperature lies exactly at the PEB temperature. After that, saturation sets in quickly [11]. Therefore, it makes sense to use a minimum PEB time of 20 min. Typical PEB temperatures are between 80 °C and 100 °C.

After PEB, there is some stress relaxation over time. Hammacher et al. [12] notice a 7.5 % drop in stress during the first 8 h after the PEB.

SU-8 Photoresist, Table 3 Absorbance (inverse of penetration depth) of unexposed SU-8 for typical wavelengths [10]

Wavelength (nm)	Absorbance (μm^{-1})
313	1.19×10^{-1}
334	5.85×10^{-2}
365	2.71×10^{-3}
405	1.41×10^{-4}
436	6.90×10^{-5}



SU-8 Photoresist, Fig. 4 Transmission of typical UV lines to the bottom of layer versus thickness of that layer, assuming no reflection of light and using the absorbances

quoted in Table 3. transmission coefficient = $e^{-\text{absorbance} \times \text{thickness}}$

After 4 days, there is an additional drop of 2 %. This indicates that it is beneficial to build a waiting period between PEB and development, diminishing adhesion loss and deformations caused by stress. Also, a ramp up and slow ramp down of the PEB temperature are beneficial. For example, a cool down step of 4 h in an oven can be employed [13].

Development

During development, the non-cross-linked parts of the SU-8 layer are removed in a solvent. Typically, propylene glycol methyl ether acetate (PGMEA) is employed, though ethyl lactate or diacetone alcohol can be used as healthier alternatives with lower vapor pressure [6]. Exposed structures will, even after postexposure baking, adsorb solvents and thus swell during development. Structures in the $10 \mu\text{m}$ range are typically saturated within a minute [9]. As the solvent evaporates again after development is over, this should not be a problem. However, swollen structures that are not cross-linked hard enough can become permanently deformed during development.

Development is dependent on agitation and on the design of the structures fabricated, and development rates as well as development times are therefore to be taken with caution. A development time of 8 min for a 50-mm-thick layer without agitation is typically sufficient. Ultrasonic agitation was found to increase the development rate by approximately eightfold. However, it can damage fragile structures. Megasonic actuation does not cause this problem. Another option is developing the wafer in an upside down position, which was found to decrease the development time by a factor 3.5 without inducing extra damage [2].

Hardbake

The hardbake is an optional extra step at the end of the SU-8 process. During the hardbake, the material is heated well above the threshold for thermally induced cross-linking. This causes, also in regions where no photoacid is present, additional cross-linking reactions to occur until saturation. Hardbaking is done for increasing the long-term stability of the material and to close small cracks that arise in some process sequences.

The additional cross-linking causes a higher chemical resistance as well: it was found impossible to remove hardbaked SU-8 structures from their substrates with solvents such as acetone or NMP (*N*-methyl-2-pyrrolidone). For this reason, a hardbake is often not advised when the SU-8 layer will have to be removed later in the fabrication process.

For the hardbake, typically a temperature between 150 °C and 250 °C is used. Following the observations of Hammacher et al. on the PEB, it can be assumed that most of the extra cross-linking is complete after 30 min. Hardbaking at 150 °C for 60 min increases internal stress to 12 MPa (tensile) in 45 % humidity air [9]. In the latter work, the built-in stress was shown to be strongly dependent on the humidity in the environment.

Removal of SU-8

The downside of the high chemical resistance of SU-8 is the difficulty to remove the material if needed. Nonhardbaked SU-8 can still be cracked up in strong solvents such as warm NMP. When this step is followed by rinsing or ultrasonification to remove the remaining flakes, a surface having a not too high topography can be clean again.

SU-8 can be cleanly removed in piranha acid, though this is aggressive toward many metals and polymers that may already be present on the wafer. Silicon and glass are not attacked at all. Also, platinum, tantalum, and gold are inert to piranha. Chromium is etched only very slowly, allowing the cleaning of contact masks in piranha after contamination with SU-8. Another application of piranha is for cleaning SU-8 off wafers for reuse.

Reactive ion etching (RIE) in oxygen plasma with small fluorine content is used as well, though traces of a nonvolatile antimony containing residue may be formed.

Typically, 4–5 % of fluorine-containing gas such as SF₆ or CF₄ is added to the reactor atmosphere to boost the etch rate significantly.

Other less widely used options are ashing in an air or oxygen atmosphere and molten salt

removal. These need a rather high temperature, 450–500 °C and 350 °C, respectively.

Finally, one can consider downstream chemical etching (DCE). DCE is an organic removal technique in which reactants produced in plasma are blown on the wafer, which is not exposed to the plasma itself. With a wafer temperature of 225 °C and a gas composition of 98 % oxygen and 2 % CF₄, an etch rate of 6.8 μm/min can be obtained [14]. Given the reasonable working temperature, the etch speed, and the low metal etch rate, DCE is likely the most promising method for industrial use.

Alternative Exposure Methods

Next to classic UV lithography, a number of other methods to selectively cross-link SU-8 exist. Though much less commonly used, they open some extra possibilities and therefore, deserve consideration. A distinction can be made between direct write methods in which structures are written pixel by pixel or voxel by voxel, and wafer-scale methods using masks which are generally much faster as writing is done in parallel over a large surface.

The photoinitiator in SU-8 can be activated by the simultaneous absorption of two 800-nm photons, instead of by a single 400-nm photon [15]. The probability of such activation rises with the square of the intensity of the exposure. This is the basis for so-called *two-photon lithography*. The nonlinear dependence makes it possible, by focusing a laser to create a high-intensity spot that is small with respect to the layer thickness, to write true 3D structures such as photonic crystals directly in a single SU-8 layer.

Proton beam writing is another direct write methods found in literature. The penetration depth of the exposing proton beam is dependent on its energy. Thus, bridge-like structures can be written in a single SU-8 layer but not the more complicated structures that are possible with two-proton lithography.

X-ray exposure has been tested as well. Becnel et al. [16] used SU-8 as a more sensitive

alternative for the PMMA normally used as photoresist in LIGA processing. The exposure time needed was in the order of 10 min, two orders of magnitude smaller than for PMMA. The very limited diffraction and low absorption of the X-rays allow for a higher resolution in thick layers than that can be offered by UV-based processing and aspect ratios greater than 100.

Finally, *inclined light UV exposure* is worth mentioning. Some researchers have experimented with exposing SU-8 with inclined UV light through a contact mask. Thus, cylinders and beams with inclined orientation can be realized. Unless countermeasures are taken, the UV light will reflect from the substrate and a mirror image of the structure will be created by the light reflecting back from the substrate surface. At intersection of two cylinders, a ball-shaped structure forms that can be used as an in-plane microlens. Others use the technique to construct microsieves (Fig. 5). Due to the high refractive index of SU-8, it is hard to attain large inclinations as explained by Snell's law: Light entering at an oblique angle from the air will be bent toward the perpendicular axis of the higher refractive index material it enters. A solution for this is to immerse wafer and mask in a higher refractive index fluid such as glycerin. When X-rays are used instead of UV, this is not important as refractive indices do not vary much.

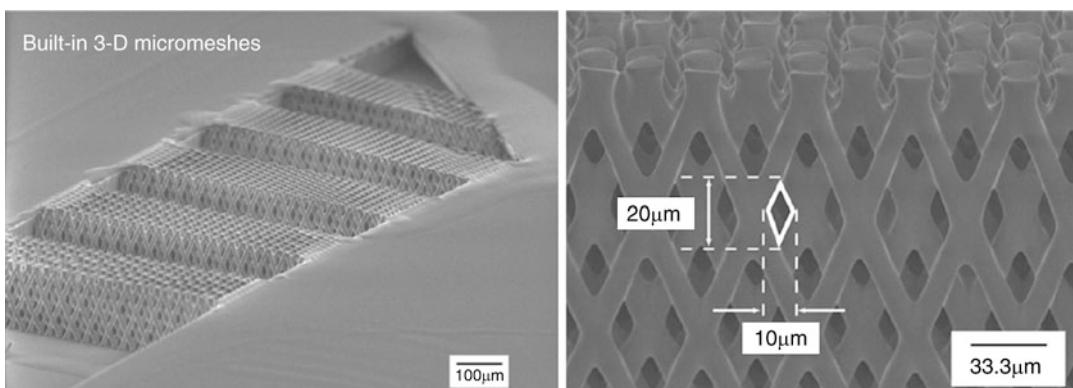
Multilayered, Freestanding Structures in SU-8

A few methods that can be used to fabricate freestanding structures such as cantilevers and microchannels out of SU-8 will be discussed here.

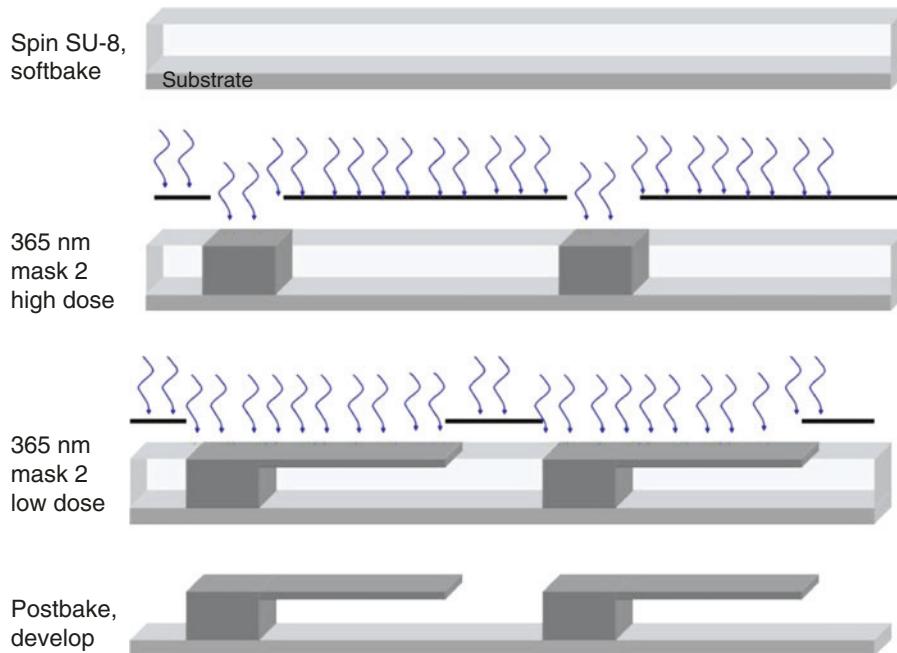
Modulation of Exposure Light

The most straightforward method is to modulate the exposure energy, that is, to use a high UV dose (i-line) to expose a single SU-8 layer from top to bottom where the anchors of the freestanding structures must come and a lower dose to define the freestanding structures (Fig. 6). However, as SU-8 is quite transparent the resulting layer thickness is very sensitive to small variations in the exposure energy and other process parameters. Also, the absorbance of SU-8 tends to increase during exposure.

Tight control of all the different parameters is required which makes it hard to achieve good tolerances. In practice, the method is only useful for fabricating thick freestanding structures with reasonable accuracy in relatively thick layers ($>250\text{ }\mu\text{m}$). Using an antireflective layer or a low-reflectivity substrate reduces the sensitivity of the process for exposure energy variations [18]. The minimum layer thickness is then reduced to $100\text{ }\mu\text{m}$.



SU-8 Photoresist, Fig. 5 Microsieve structure in microchannel, fabricated in SU-8 using inclined UV exposure ([17], used with permission) *Left*: overview. *Right*: detail of sieve



SU-8 Photoresist, Fig. 6 Principle of the creation of freestanding structures by exposure time control

The above process can be improved by using different wavelengths in the two UV exposure steps. By using a wavelength that is quickly absorbed for the second exposure step, thinner freestanding structures can be made with increased repeatability. It is straightforward to use the 313-nm wavelength for this as it is already present in the spectrum of standard exposure tools. The practical thickness range that can be reached this way is between 6 and 25 μm (Fig. 7). Due to uneven exposure between the top and bottom parts of the freestanding structure, single-clamped beams were observed to bend down [2].

Buried Mask Process

In this process two SU-8 layers are used, as illustrated in Fig. 8. Thus, a larger thickness range and a better control of layer thicknesses are obtained at the cost of a higher process complexity. First, the lower layer is processed as usual up to the PEB step. Then, a UV-blocking layer is deposited. The purpose of this layer is to shield unexposed parts of the lower layer from the UV light used to expose the upper layer, later in the process. Then, the upper layer is spun on and processed.

The two layers are developed together at the end of the process sequence. Halfway in the development step, the UV-blocking layer must be removed when it is not dissolving in the normal SU-8 developer, for example, when a metal UV-blocking layer is used. It is important to deposit the UV-blocking layer such that no excess UV or heat is produced, which would cross-link SU-8 undesirably. Also, temperature must be kept at a minimum when processing the second SU-8 layer to avoid wrinkling [2]. Figure 9 shows some example structures fabricated with the process.

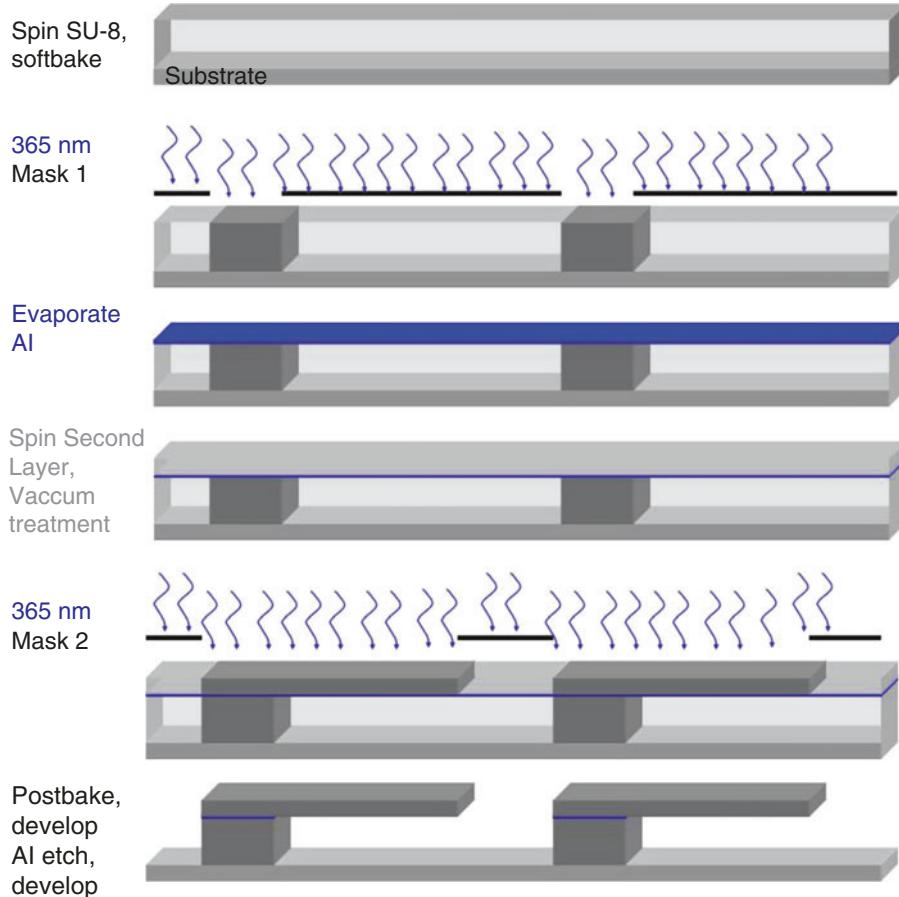
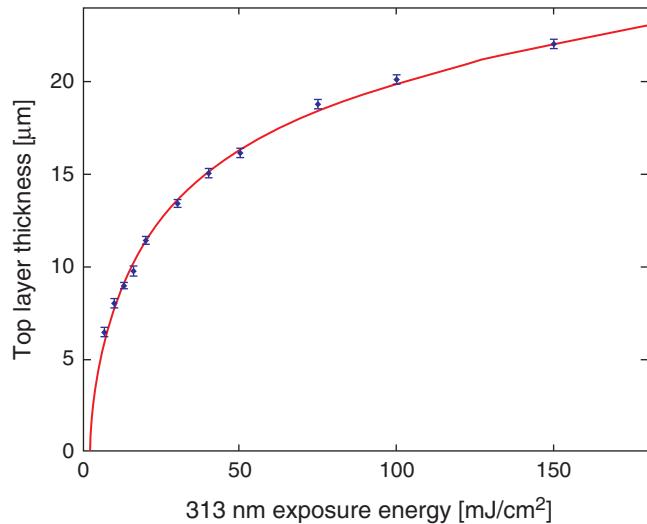
Evaporation by resistive heating of a low melting temperature metal such as aluminum, zinc, or magnesium is a suitable method. Another author uses a metal layer transfer process based on a gold-covered silicone stamp [19].

Sacrificial Layer-Based Processes

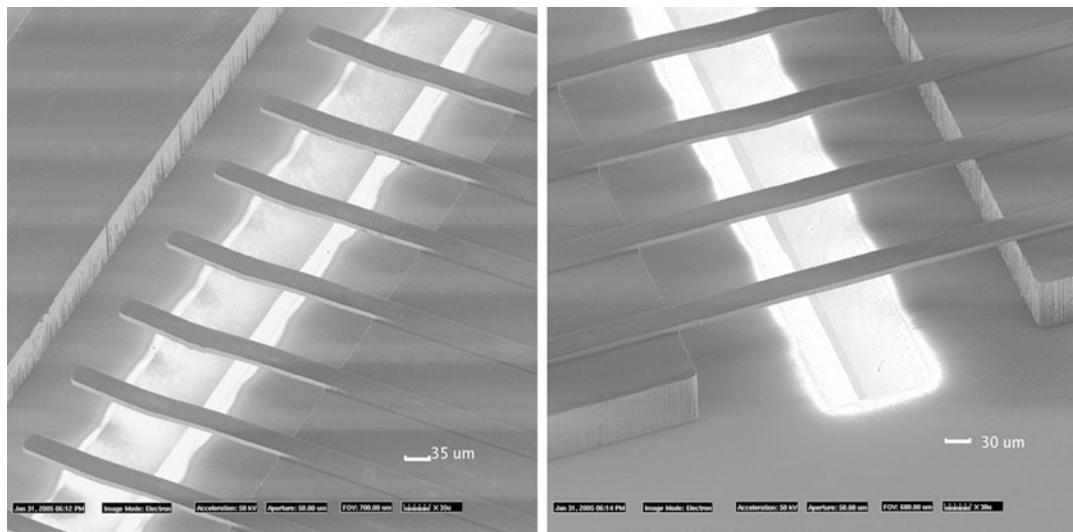
A third major category of processes used to fabricate freestanding SU-8 structures is based on the deposition of a sacrificial layer in another material. After the SU-8 layer is processed on top of the sacrificial layer the sacrificial layer is dissolved, creating freestanding SU-8 structures.

SU-8 Photoresist,

Fig. 7 UV exposure energy at 313-nm wavelength versus resulting freestanding layer thickness showing LSE fit. Data points from measurement with (tiny) error bars showing standard deviation are plotted as well [2]



SU-8 Photoresist, Fig. 8 Buried mask process



SU-8 Photoresist, Fig. 9 Single- and double-clamped freestanding beams, fabricated with the buried mask method, lying over a KOH-etched U-groove

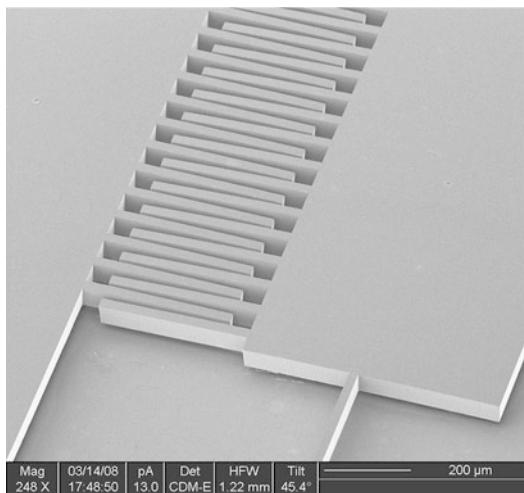
The sacrificial layer can be patterned before SU-8 deposition to create anchors defining where the SU-8 structures that will be formed in the next step will be attached to the substrate. When no anchor sites are defined either a well-timed etch stop can define anchors or loose SU-8 components are created.

The main property a sacrificial layer must have is that there must be a way to remove it without damaging SU-8. Reversely, the solvent in SU-8 must not dissolve the sacrificial layer to prevent layer intermixing and unpredictable results. Another point of interest is the adhesion of SU-8 on the sacrificial layer, which must be sufficient to allow it to survive the processing. The other properties that discern the different sacrificial layers discussed here are their layer thickness range, processing cost and time, and achievable aspect ratio.

A metal sacrificial layer is a first option. Many metals common to micromachining, such as aluminum and copper, can be etched chemically with perfect selectivity with respect to SU-8. Also, the adhesion of SU-8 to metals is generally satisfying (Table 2). There are some disadvantages to the use of metals, though. The most straightforward way to apply a metal layer is by a thin film technique such as sputtering. Due to stress and

relatively slow speed, these processes are limited to a thickness of a few microns. Furthermore, the attainable aspect ratio of the lower layer is determined by the etching process of the metal and is typically around one unless reactive ion etching is used. Both disadvantages can be overcome at the cost of an extra process step by the use of a sacrificial layer that is electroplated in a photore sist mold. Still, when using a metal sacrificial layer, the etch process of the metal may pose constraints on other metal layers that might be present and attacked as well by the release etch. A final point is that metal deposition methods are never self-planarizing, limiting the use on uneven surfaces.

A second candidate is polyimide sacrificial layers. Polyimides resist the solvent in SU-8 very well once properly cross-linked. However, their removal is a problem. Polydimethylglutarimide (PMGI), on the other hand, is etchable in alkaline solutions such as positive photoresist developer while still being resistant to solvents. Thus, it can be removed selectively with respect to cross-linked SU-8 and is easily patternable. Disadvantages are the limited lower layer thickness and the fact that the weak alkaline solution will still etch aluminum and, after an incubation time in which native oxide is removed, even (slowly)



SU-8 Photoresist, Fig. 10 Comb actuator structure fabricated in SU-8 using a PMGI sacrificial layer

silicon. In Ref. [2], comb actuator structures were fabricated based on this process (Fig. 10).

Positive photoresists can also be considered as sacrificial layer. They are inexpensive, can be quickly applied, and are available in a thickness range from below 1 μm to about 100 μm . Therefore, they would be ideally suited as sacrificial layer for a wide range of applications. The main problem hindering their applicability as sacrificial layer for SU-8 is that they dissolve in the solvent (cyclopentanone or GBL) present in SU-8. There are two ways to prevent this. The first is to deposit a thin metal layer on top of the positive resist. The second one is to strongly hardbake the positive photoresist. However, in the latter case resist reflow typically occurs, which severely limits the attainable aspect ratio of the sacrificial layer.

Other sacrificial polymers used by workers in the field are polystyrene (PS) and PMMA. Toluene can be used as a solvent for spincoating and release, as cured SU-8 is resistant to several hours of immersion in toluene.

It is possible to use RIE to pattern the polymers or to include photosensitizers in the PMMA, such as is done for deep-UV photoresists.

A final interesting type of sacrificial polymers is polymers that decompose into volatile components cleanly at a temperature SU-8 can withstand. Examples are polypropylene carbonate

(PPC) and polyethylene carbonate (PEC) [20]. Dissolved in NMP (*N*-Methyl-2-pyrrolidone), they can be deposited by spin coating. Structuring is possible using oxygen plasma. PEC starts to decompose thermally around 220 °C in nitrogen at atmospheric pressure. For PPC this is around 240 °C. The most remarkable property of this process is that the volatile decomposition products can diffuse throughout SU-8. Thus, the sacrificial layer can even be removed from closed cavities, and long channels can be cleared in a time independent of channel length. For a 30-mm-thick SU-8 layer, the sacrificial layer removal process time is about 10 min. For very thick upper layers, this time might be considerably longer.

Layer Transfer Processes

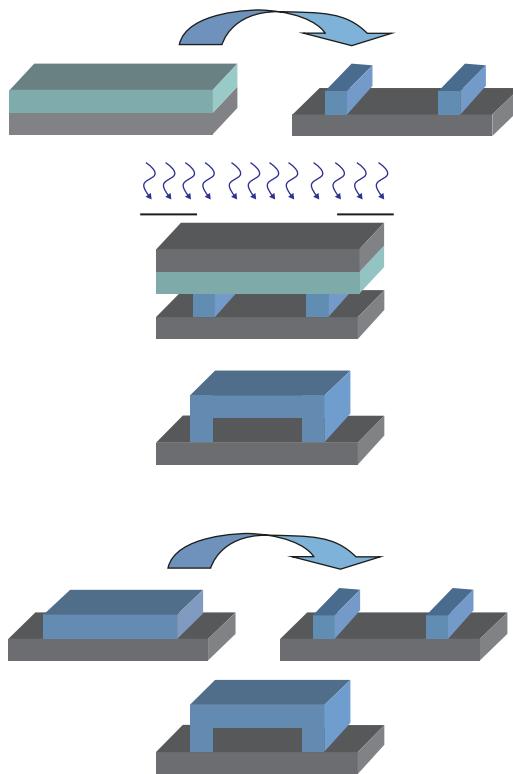
A final major approach to fabricating freestanding structures in SU-8 is to fabricate a single SU-8 layer on a carrier substrate and then transfer that layer to a second substrate on which another SU-8 layer is already present.

This has the advantage that closed cavities or long channels can be fabricated. As the sealed cavities may contain released micromechanical structures, these processes can be used for capping purposes as well.

The different variations of the processes can be further categorized according to the process step after which the layer transfer and optionally the removal of the carrier substrate is done (Fig. 11).

When the transfer is done after softbaking, typically by applying limited heat and pressure, the material is still thermoplastic. Therefore, in order to achieve good control over the shape of the fabricated structure care must be taken to select process parameters such that little flowing of the non-cross-linked layer occurs. A typical sign of too much reflow is the clogging of channels that were sealed by the process and rounded corners.

Another way to limit channel clogging is to use a micron-thin softbaked SU-8 layer to bond two substrates together. Bonding is done at 2 bars in a vacuum bonder [21]. However, the advantage that the upper layer can be readily patterned, for example, integrating vertical microfluidic connectors, is thereby lost.



SU-8 Photoresist, Fig. 11 Layer transfer processes. *Top:* cross-linking after transfer. Sometimes the carrier wafer is not removed, and SU-8 serves as an adhesive between the two substrates. *Bottom:* transfer of partially cross-linked layers

As the transferred SU-8 layer still must be cross-linked, either the carrier wafer must be UV transparent or it must be removed before exposure. After exposure and PEB, the carrier wafer must be removed in any case, unless the SU-8 layer is used as a simple adhesive. The removal is not straightforward with common hard substrates covered with a sacrificial material as this involves under etching of that sacrificial material for several centimeters. A better way is to use a foil as carrier, which can be laminated on with relatively low pressure, avoiding air entrapment, and channel clogging. Furthermore, a foil can be removed by simply peeling it off.

A polyimide foil should be peeled off before exposure as it is not UV transparent. Teflon foil, on the other hand, is transparent to UV.

Due to the rather rough nature of the carrier wafer removal process, it is best used when the top layer contains relatively large features. Examples are device capping and the sealing of microchannels.

Deformation of the transferred layer can be avoided altogether by cross-linking it first before transferring it. The transferred layer can then also be developed before transfer.

To make sure sufficient reactivity remains available to enable thermocompression bonding, the exposure energy and PEB must be scaled down. Arroyo et al. [22] determined the optimal parameters as 140 mJ/cm^2 for i-line exposure followed by a PEB of 4' at 85°C . Due to the decreased deformability of the SU-8, a much higher bonding force is needed during thermocompression bonding: good results (95 % yield) were achieved with a pressure 3.25 bar applied at a temperature of 88°C for 12 min.

Furthermore, it is of course essential to have good thickness uniformity. The edge bead must certainly be removed, be it mechanically, chemically, or by designing the mask such that the rim of the wafer is not exposed. In this case too, it can be advantageous to use a Teflon or polyimide foil as carrier for the second layer as it can be easily pulled off.

Electrically Active Structures in SU-8

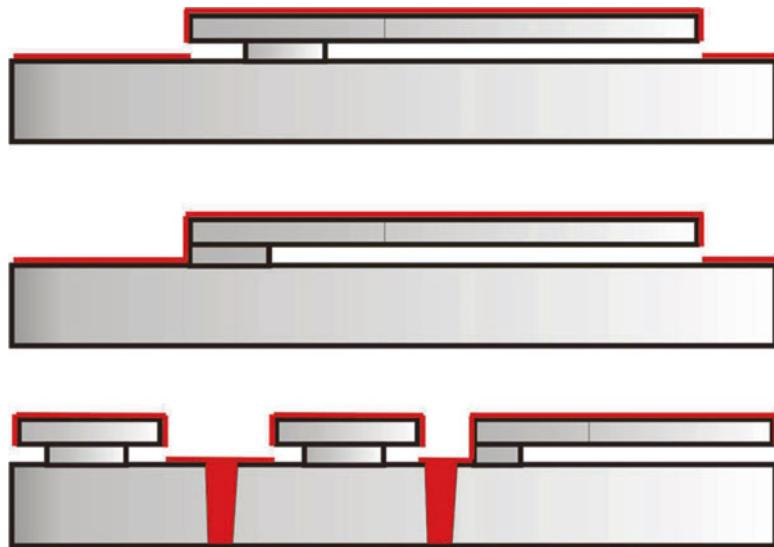
In this paragraph, a few strategies will be discussed to render SU-8 conductive, which is mandatory for many actuation and sensing applications such as comb drives and capacitive accelerometers.

Conductive polymers such as polyaniline (PANI) can be added to the SU-8 resin. However, it was shown that high PANI quantities (up to 50 %) are required in order to achieve useful conductance, a point at which many SU-8 properties such as adhesion degrade.

A different route to create conductive SU-8 is to use a silver nanoparticle-based filler. A silver content of 6 vol% is enough to reach to the percolation threshold, and to create a conductivity of over 10 S cm^{-1} , which can go up a factor 100 with

SU-8 Photoresist,

Fig. 12 Metallization of SU-8 by sputter deposition of a thin metal layer (red). *Top:* isolated freestanding structure. *Middle:* connection to substrate. *Bottom:* isolation of bonding pad (*left*)



a higher silver content [23]. Such blends are commercially available. However, because of light reflected by the filler, the available aspect ratio and achievable thickness are limited to about 1 and 10 μm , respectively. Carbon nanotubes have been explored as filler material as well. However, conductivities are typically quite limited.

A straightforward method to achieve conductive SU-8 structures is to sputter deposit a thin metal layer. As during sputtering material gets deposited on top of as well as on sidewalls of structures, the mask layout must be designed such as not to cause short circuits, for example, by incorporating overhanging structures such as illustrated in Fig. 12 [2].

Conclusion

Concluding this oversight of SU-8 photoepoxy processing, there is little doubt that this material will find – and indeed already has found – its way into a wide range of applications that benefit from the unique combination of low cost wafer-scale fabrication and the high aspect ratios and very high layer thicknesses attainable. Examples of those are the fabrication of microchannels, labs-on-a-chip, integrated optics, ink jet heads, and electroplating molds for fabricating metal

structures. For the latter, electroplating in SU-8 is a low cost alternative for the LIGA process, provided minimum features are several microns wide and more relaxed tolerances can be acceptable. As SU-8 is a low-temperature process, it even shows potential for the fabrication of sensors directly on CMOS wafers.

Cross-References

- ▶ [Lab-on-a-Chip for Studies in *C. elegans*](#)
- ▶ [Microfluidic Whole-Cell Biosensor](#)
- ▶ [Plating](#)
- ▶ [Stereolithography](#)

References

1. Zhang, J., Tan, K.L., Hong, G.D., Yang, L.J., Gong, H. Q.: Polymerization optimization of SU-8 photoresist and its applications in microfluidic systems and MEMS. *J. Micromech. Microeng.* **11**, 20–26 (2002)
2. Ceyssens, F.: Micromachining in polymers and glass: process development and applications. Ph.D. thesis, KULeuven, Leuven, Belgium (2009)
3. Schoeberle, B., Wendlandt, M., Hierold, C.: Long-term creep behavior of SU-8 membranes: application of the time-stress superposition principle to determine the master creep compliance curve. *Sens. Actuators A Phys.* **142**, 242–249 (2008)
4. Chung, C.K., Hong, Y.Z.: Surface modification of SU8 photoresist for shrinkage improvement in a

- monolithic MEMS microstructure. *J. Micromech. Microeng.* **17**, 207–212 (2007)
5. Bohl, B., Steger, R., Zengerle, R., Koltav, P.: Multi-layer SU-8 lift-off technology for microfluidic devices. *J. Micromech. Microeng.* **15**, 1125–1130 (2005)
 6. MicroChem: SU-8 datasheet and adhesion results-shear analysis. www.microchem.com (2007). Accessed 18 Nov 2011
 7. Teh, W.H., Dürig, U., Drechsler, U., Smith, C.G., Gntherodt, H.-J.: Effect of low numerical-aperture femtosecond two-photon absorption on SU-8 resist for ultrahigh-aspect-ratio microstereolithography. *J. Appl. Phys.* **97**, 4095 (2005)
 8. Becnel, C., Desta, Y., Kelly, K.: Ultra-deep x-ray lithography of densely packed SU-8 features: I. An SU-8 casting procedure to obtain uniform solvent content with accompanying experimental results. *J. Micromech. Microeng.* **15**, 1242–1248 (2005)
 9. Wouters, K., Robert, P.R.: Diffusing and swelling in SU-8: insight in material properties and processing. *J. Micromech. Microeng.* **20**, 095013 (2010)
 10. Reznikova, E.F., Mohr, J., Hein, H.: Deep photolithography characterization of SU-8 resist layers. *J. Microsyst. Technol.* **11**, 282–291 (2005)
 11. Feng, R., Farris, R.: Influence of processing conditions on the thermal and mechanical properties of SU8 negative photoresist coatings. *J. Micromech. Microeng.* **13**, 80–88 (2003)
 12. Hammacher, J., Fuelle, A., Flaemig, J., Saupe, J., Loechel, B., Grimm, J.: Stress engineering and mechanical properties of SU-8-layers for mechanical applications. *J. Microsyst. Technol.* **14**, 1515–1523 (2007)
 13. Williams, J.D., Wang, W.: Using megasonic development of SU-8 to yield ultrahigh aspect ratio microstructures with UV lithography. *J. Microsyst. Technol.* **10**, 694–698 (2004)
 14. Dentinger, P.M., Miles, C., Goods, S.H.: Removal of SU-8 photoresist for thick film applications. *Microelectron. Eng.* **61–62**, 993–1000 (2002)
 15. Witzgall, G., Vrijen, R., Yablonovitch, E., Doan, V., Schwartz, B.J.: Single-shot two-photon exposure of commercial photoresist for the production of threedimensional structures. *Opt. Lett.* **23**, 1745 (1998)
 16. Becnel, C., Desta, Y., Kelly, K.: Ultra-deep x-ray lithography of densely packed SU-8 features: II. Process performance as a function of dose, feature height and post exposure bake temperature. *J. Micromech. Microeng.* **15**, 1249–1259 (2005)
 17. Sato, H., Matsumura, H., Keino, S., Shoji, S.: An all SU-8 microfluidic chip with built-in 3D fine microstructures. *J. Micromech. Microeng.* **16**, 2318–2322 (2006)
 18. Chuang, Y., Tseng, F., Cheng, J., Lin, W.: A novel fabrication method of embedded micro-channels by using SU-8 thick-film photoresists. *Sens. Actuators A Phys.* **103**, 64–69 (2003)
 19. del Campo, A., Greiner, C.: SU-8: a photoresist for high-aspect-ratio and 3D submicron lithography. *J. Micromech. Microeng.* **17**, R81–R95 (2007)
 20. Metz, S., Jiguet, S., Bertsch, A., Renaud, P.: Polyimide and SU-8 microfluidic devices manufactured by heat-depolymerizable sacrificial material technique. *Lab Chip* **4**, 114–120 (2004)
 21. Carlier, J., Arscott, S., Thomy, V., Fourrier, J.C., Caron, F., Camart, J.C., Druon, C., Tabourier, P.: Integrated microfluidics based on multi-layered SU-8 for mass spectrometry analysis. *J. Micromech. Microeng.* **14**, 619–624 (2004)
 22. Arroyo, M.T., Fernández, L.J., Agirregabiria, M., Ibañez, N., Aurrekoetxea, J., Blanco, F.J.: Novel all-polymer microfluidic devices monolithically integrated within metallic electrodes for SDS-CGE of proteins. *J. Micromech. Microeng.* **17**, 1289–1298 (2007)
 23. Jiguet, S., Bertsch, A., Hofmann, H., Renaud, P.: Conductive SU8-silver composite photopolymer. In: Proceedings of Micro Electro Mechanical Systems, Maastricht, The Netherlands. pp. 125–128 (2004)
 24. Yang, R., Wang, W.: A numerical and experimental study on gap compensation and wavelength selection in UV-lithography of ultra-high aspect ratio SU-8 microstructures. *Sens. Actuators B* **110**, 279–288 (2005)

S

Sub-retinal Implant

- Artificial Retina: Focus on Clinical and Fabrication Considerations

Subwavelength

- Moth-Eye Antireflective Structures

Sub-wavelength Waveguiding

- Light Localization for Nano-optical Devices

Superelasticity and the Shape Memory Effect

Xiaodong Han¹, Shengcheng Mao¹ and Ze Zhang^{1,2}

¹Institute of Microstructure and Property of Advanced Materials, Beijing University of Technology, Chaoyang District, Beijing, People's Republic of China

²State Key Laboratory of Silicon Materials and Department of Materials Science and Engineering, Zhejiang University, Hangzhou, China

Synonyms

Hyperelasticity; Pseudoelasticity; Theoretical elasticity; Ultralarge strain elasticity

Definition

Superelasticity, or pseudoelasticity, is a unique property of shape memory alloys (SMAs), wherein up to 13 % deformation strain can be sustained and the material can recover its original shape after removing the stress. The shape memory effect occurs in SMA and is defined as when a material can remember its original shape upon heating or cooling.

Overview

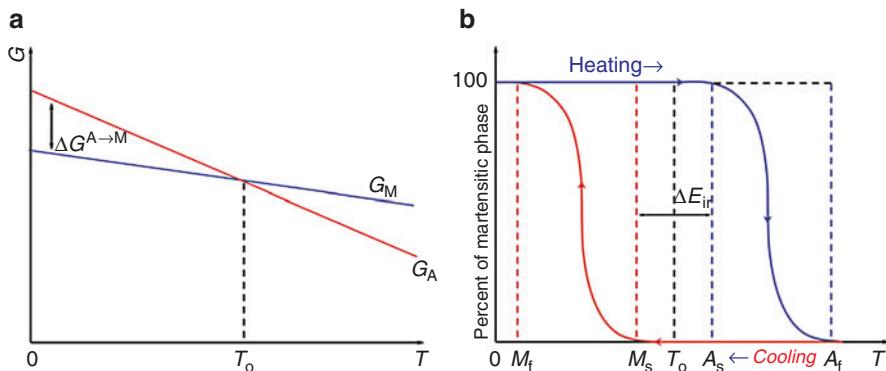
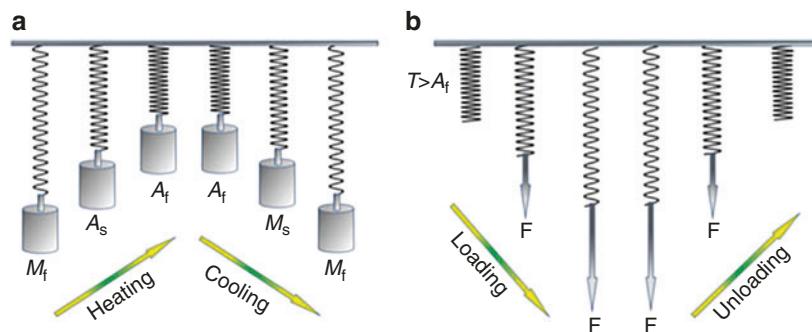
Superelasticity (SE), also called pseudoelasticity, normally refers to a phenomenon observed in shape memory alloys (SMAs), a kind of metal that can remember its original shape after heating or cooling [1–3]. SE is different from true elasticity in bulk metals, which is a reflection of the interatomic spacing variation within a material (Hooke's law). Regular elasticity in bulk metals is normally less than 0.5 %, whereas the elasticity of shape memory alloys can be up to 13 % in Fe-based SMA [4]. This SE occurs through

reversible martensitic transformations from austenite to martensite and vice versa. Upon external mechanical loading or temperature change (or a mixture of the two), the SMA deforms by reversible martensitic phase transformations rather than by irreversible plastic dislocation glide. The martensitic transformation occurs through shear on an invariable plane (habit plane) along the shear direction. SE occurs when the metals are deformed in the austenite state, and the deformation shape is recoverable when the applied load is removed.

Large strain elasticity (LSE) can be achieved in nanomaterials due to their defect-free structure, particularly in nanowires. The true elasticity of metallic nanowires can even approach the theoretical elastic limit (8 %) [5], and this is typically associated to extremely high strength [6]. In theory and in computer simulations, some metallic nanowires, such as Cu nanowires, can reversibly deform up to 50 % by a combination of large strain elasticity and pseudoelasticity [7]. Some of these exceptional mechanical properties have already been experimentally demonstrated in metallic nanowires.

The *shape memory effect* (SME) also relates to SMAs, and it refers to the phenomenon where a metallic specimen changes shape in response to temperature changes [1–3]. Figure 1 shows a sketch comparing SME and SE. A deformed SMA can return to its original high-temperature shape upon heating or go back to the low-temperature shape upon cooling (see Fig. 1a). If a SMA is deformed by mechanical loading at a constant temperature higher than a critical temperature (the austenite finish temperature, defined below), the initial shape can be recovered by unloading, through a reverse martensitic transformation (Fig. 1b). The SME was first discovered by A. Olander in Au-Cd alloy in 1932. Investigations into SMA and their applications have been further promoted since the discovery of the SME in the near-equatomic NiTi alloys, known as "Nitinol." The SME has also been found in other metallic alloys such as Cu-Al-Ni, Cu-Zn, Fe-Mn-Si, Ni-Mn-Ga, and Ti-Ni-Hf. Because of these unique properties,

Superelasticity and the Shape Memory Effect, Fig. 1 A schematic illustration of (a) the shape memory effect and (b) superelasticity



Superelasticity and the Shape Memory Effect, Fig. 2 Schematic illustration of the evolutions of (a) the Gibbs free energies of the martensite and austenite phases and (b) the martensitic fraction with temperature

SMAs have been used for a variety of applications, including orthodontic wire, stents, microactuators, pipe coupling, eyeglass frames, microelectromechanical systems, and a variety of biomedical devices [8, 9]. *Superelasticity and the shape memory effect* are correlated with thermoelastic and stress/strain-induced martensitic transformations. In the following, we introduce some basic concepts of *martensitic transformations* in NiTi SMAs.

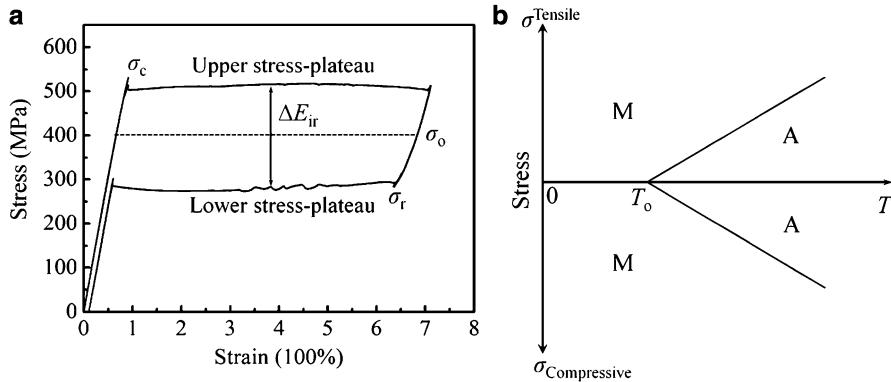
Martensitic Transformation

Thermodynamic Consideration of the Thermoelastic Martensitic Transformation

The martensitic transformation is a transformation between austenite (stable at high temperature) and martensite (stable at low temperature), and it is usually treated as a diffusionless first-order phase

transformation. The Gibbs free energy, defined as $\Delta G = \Delta H - T\Delta S$ (where ΔH and $T\Delta S$ are the changes in chemical enthalpy and entropy), is the pertinent chemical potential of a system, which is at the minimum when reaches equilibrium at constant pressure and temperature. The martensite and austenite Gibbs free energies (ΔG^M and ΔG^A) gradually decrease, following different slopes, with increasing temperature (T), as illustrated in Fig. 2a. At temperature T_0 , the Gibbs free energies of the two phases reach a thermodynamic equilibrium ($\Delta G^M = \Delta G^A$). To drive the forward or reverse martensitic transformation, the energy gap between the two phases ($\Delta G^{A \rightarrow M}$ or $\Delta G^{M \rightarrow A}$) needs to be overcome by either temperature or mechanical loading, as expressed in Eqs. 1 and 2: [10].

$$\begin{aligned}\Delta G^{A \rightarrow M} &= \Delta H^{A \rightarrow M} - T\Delta S^{A \rightarrow M} - \Delta E_{mech} \\ &= \Delta G^{A \rightarrow M} - T\Delta S^{A \rightarrow M} - \frac{\sigma \epsilon}{\rho}\end{aligned}\quad (1)$$



Superelasticity and the Shape Memory Effect,

Fig. 3 (a) A typical stress–strain curve of NiTi SMA showing superelasticity. (b) Schematic illustration of the

stress–temperature relationship of an SMA. The stress can be both tensile and compressive stress

$$\Delta G^{M \rightarrow A} = \Delta H^{M \rightarrow A} - T \Delta S^{M \rightarrow A} + \frac{\sigma \varepsilon}{\rho} \quad (2)$$

where ρ , σ , and ε are the density of materials, the external stress, and strain, respectively.

Figure 2b shows schematically the evolution of the martensite volume fraction with temperature during forward and reverse transformations. There are four characteristic temperatures defining a thermoelectric martensitic transformation: (1) The martensite start temperature, M_s , at which martensite starts to nucleate; (2) the martensite finish temperature, M_f , with $M_f < M_s$, at which the transformation has completed; (3) the austenite start temperature, A_s , at which the system starts moving from martensite to austenite upon heating; and (4) the austenite finish temperature, A_f , with $A_f > A_s$, at which the martensitic phase has fully transformed into the austenite phase.

Aside from the thermally induced martensitic transformation, martensite can also be obtained by applying external stress. Figure 3a shows a typical stress–strain curve of NiTi SMA showing superelasticity. There are two stress plateaus, defined as the upper stress plateau and the lower stress plateau, corresponding to the forward and reverse martensitic transformation, as indicated in Fig. 3a. The martensite starts to nucleate after reaching a critical transformation stress, σ_c . The martensitic transformation then proceeds at nearly constant stress upon further straining (the physical mechanisms are described below). As shown in

Fig. 3a, the upper plateau stress is lower than the critical transformation stress. As a result, the growth energy is smaller than that of the nucleation energy of martensite. The austenite nearly fully transforms to martensite at the end of the upper stress plateau. Upon unloading to a critical reverse stress, σ_r , martensite starts transforming back to austenite. The martensite phase is nearly fully reverted to the austenite phase at the end of the lower stress plateau. According to Eq. 1, higher stress is required to trigger martensite transformations as the ambient temperature increases. The stress–temperature relationship in stress-induced martensitic transformations is schematically illustrated in Fig. 3b.

As shown in Figs. 2b and 3a, the forward and reverse deformation paths do not overlap, i.e., temperature and stress hysteresis exist during thermo- and stress-induced martensitic transformation. The austenite start temperature is higher than the martensite start temperature ($A_s > M_s$, $A_s > T_0$ and $M_s < T_0$), and the unloading stress plateau is lower than the loading stress plateau. This is because the generation of irreversible energies (ΔE^{ir}), including the frictional energy of the martensite–austenite phase interface propagation, the acoustic emission, and the production and motion of dislocations, is ineluctable during the martensitic transformation [11]. Therefore, additional chemical or mechanical energy is required to promote the forward and reverse martensitic transformations.

Crystallography of Stress-Induced Martensitic Transformation

Critical Transformation Stress

Two important characteristics, the critical transformation stress and transformation strain, define the stress-induced martensitic transformation. It has been suggested that the selection of a martensitic variant under external stress (uniaxial tension and compression) comes from satisfying Schmid's law, as illustrated in Fig. 4 [12]. The stress component along the shear direction in the habit plane can be expressed as:

$$\begin{aligned}\tau_s \frac{F_s}{A_s} &= \frac{F \cos \lambda}{A_o / \cos \varphi} = \frac{F}{A_o} \cos \lambda \cdot \cos \varphi \\ &= \sigma \cos \lambda \cdot \cos \varphi\end{aligned}\quad (3)$$

where $m = \cos \lambda \cdot \cos \varphi$ is defined as the Schmid factor, λ is the angle between the loading axis and the shear direction, and φ is the angle between the loading axis and the normal of the habit plane.

According to Eq. 3, the variant with the highest Schmid factor and, consequently, the highest resolved shear stress will be triggered in one grain, which shows that the critical transformation

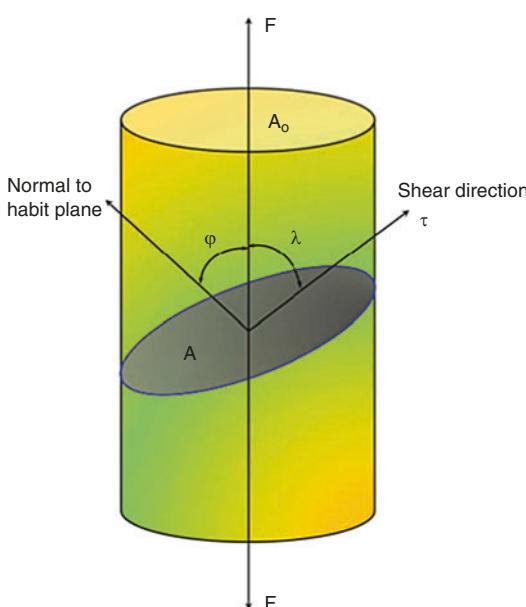


Fig. 4 Schematic illustration of Schmid's law

stress of martensite is strongly related to the grain/crystal orientation.

Maximum Recoverable Strain

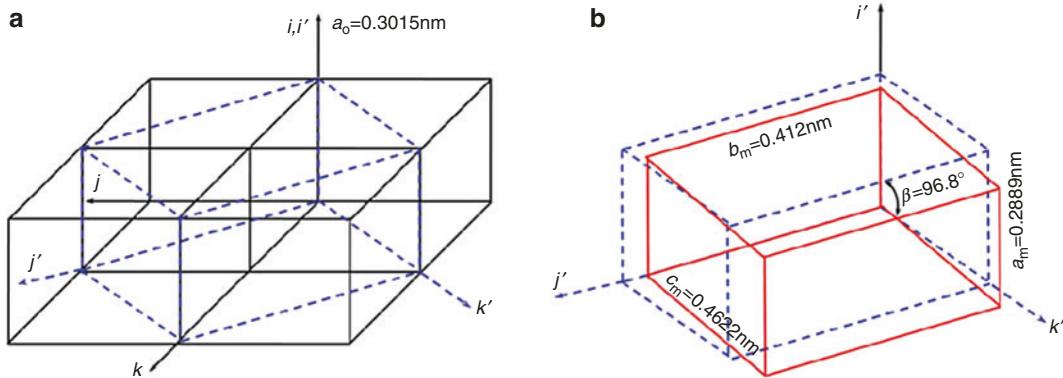
The two phases of NiTi alloys, austenite and martensite, are known to have cubic (B2) and monoclinic (B19') structures. Figure 5 shows a schematic illustration of the lattice distortion of the B2–B19' transformation in NiTi SMA. A martensitic unit cell is identified by the dashed line within the austenite lattice, as shown in Fig. 5a. The embryo may be transformed to the martensitic unit cell by a simple "bain distortion," as indicated in Fig. 5b, which can be expressed as an expansion in c , contractions in the a and b directions, and a change of the beta angle from 90° to 96.8°.

Such a transformation can happen in 12 equivalent lattice correspondence martensitic variants (LCMVs) in NiTi [13]. The lattice distortion shown in Fig. 5 corresponds to LCMV #1 with lattice correspondences of $[001]_m-[001]_a$, $[010]_m-[011]_a$, and $[001]_m-[0-11]_a$. The lattice deformation matrix M' in the coordinates of the martensite (i', j', k') can be expressed using the lattice constants of the parent and the product phases. For the martensitic transformation in near-equiatomic NiTi, the lattice constant of the austenite is $a_o = 0.3015$ nm and those of the martensite are $a_m = 0.2889$ nm, $b_m = 0.412$ nm, $c_m = 0.4622$ nm, and $\beta = 96.8^\circ$ [14]; thus:

$$M' = \begin{bmatrix} a_m & 0 & c_m \cos \beta \\ a_o & \frac{b_m}{\sqrt{2}a_o} & 0 \\ 0 & \frac{\sqrt{2}a_o}{b_m} & 0 \\ 0 & 0 & \frac{c_m \sin \beta}{\sqrt{2}a_o} \\ 0.9582 & 0 & 0.1283 \\ 0 & 0.9663 & 0 \\ 0 & 0 & 1.0763 \end{bmatrix} \quad (4)$$

The lattice deformation matrix M in the coordinates of austenite (i, j, k) can then be transformed from M' with a coordinate transformation matrix R from the martensite to the austenite via:

$$M = RM'R^T \quad (5)$$



Superelasticity and the Shape Memory Effect,
Fig. 5 Schematic illustration of the lattice distortion of the B2–B19' martensitic transformation in NiTi: (a) atomic coordinate systems, with \$(i, j, k)\$ representing the reference

where \$R^T\$ is the transpose of \$R\$. Using the lattice correspondence of austenite and martensite, the coordinate transformation matrix \$R\$ for the variant shown in Fig. 5 is expressed as:

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \quad (6)$$

The lattice deformation matrix \$M\$ is then calculated to be:

$$M = RM'R^T = \begin{bmatrix} \alpha & -\beta & \beta \\ 0 & \omega & -\gamma \\ 0 & -\gamma & \omega \end{bmatrix} \quad (7)$$

where \$\alpha = 0.9582\$, \$\beta = 0.0907\$, \$\omega = 1.0213\$, and \$\gamma = 0.0550\$.

With this, a vector \$\mathbf{x}\$ in the austenite is transformed to \$\mathbf{x}'\$ upon the martensitic transformation by the following equation:

$$\mathbf{x}' = M\mathbf{x} \quad (8)$$

Consequently, the transformation strain can be calculated by:

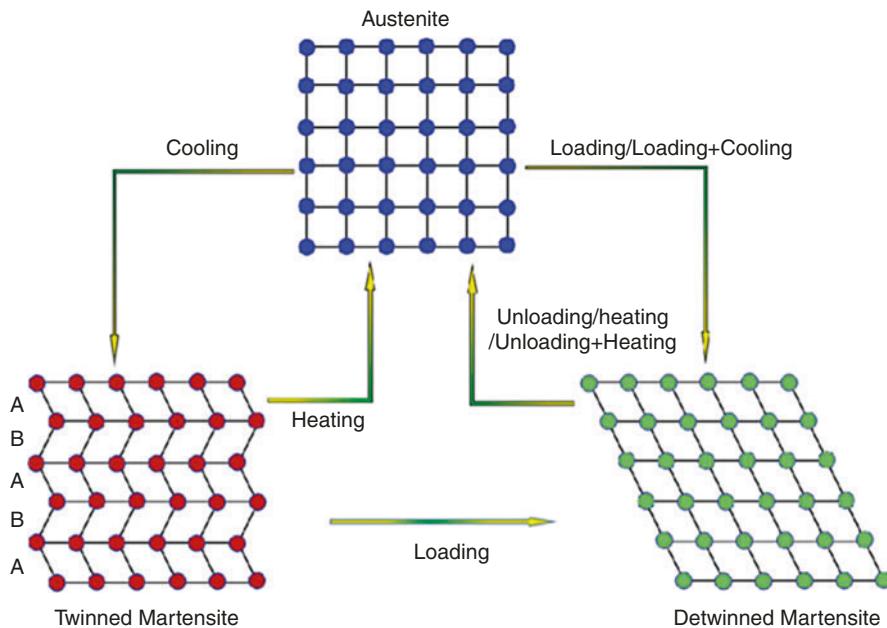
$$\varepsilon = \frac{|\mathbf{x}'| - |\mathbf{x}|}{|\mathbf{x}|} \quad (9)$$

frame in the austenite and \$(i', j', k')\$ representing the reference frame in the martensite, (b) lattice distortion from B2 (dashed lattice) to B19' (solid lattice)

According to thermodynamic principles, the variant with the maximum strain in the direction of the applied load produces the maximum driving force for the stress-induced martensitic (SIM) transformation [15] and is triggered to form first.

Mechanisms of the Shape Memory Effect and Superelasticity

The unique properties, superelasticity and the shape memory effect, of SMA can be expressed by a simplified 2D geometric depiction, as illustrated in Fig. 6. Upon cooling to a temperature \$T < M_f\$, two martensitic variants, defined as A and B, with the same crystal structure but different orientations were cooperatively triggered to accommodate and minimize the martensitic transformation strain. The interface between the thermally induced martensite, composed of the matrix and its twin, and the austenite is the habit plane. The martensite variant at the habit plane is then referred to as the habit plane variant. Under external stress (\$T < M_f\$), martensitic variant A grows at the expense of variant B by the movement of the interface of variants A and B. The deformation process is referred to as detwinning or reorientation of the martensites. The twinned or detwinned martensitic variants can fully transform back to austenite when heating the specimen



Superelasticity and the Shape Memory Effect, Fig. 6 Schematic illustration of the martensitic structure upon loading, unloading, heating, and cooling

to a temperature above A_f corresponding to the recovery of the macroscopic SMA to its original shape. There are two types of SME: one-way and two-way. The difference is whether the SMA remembers its low-temperature shape. The critical stress to drive the reorientation of martensite is known to be very small (about 100 MPa); therefore, it is very easy to set the shape of SMA at low temperatures ($T < M_f$). For the one-way SME, the shape of SMA will remain stable until heated above the reverse transformation start temperature (A_s) and reverse to its original when the temperature is higher than A_f . Upon the second cooling to a temperature lower than M_f , no macroscopic shape change happens, indicating that only the high-temperature shape is “remembered.”

The two-way SME is an effect where both the low-temperature and high-temperature shapes can be “remembered.” Thus, when cooling the SMA to a temperature $T < M_f$, the high-temperature shape will change to the low-temperature shape; likewise, the low-temperature shape will transform back to the high-temperature shape on heating to $T > A_f$. The shape variations between the two shapes are stress free. It is an intrinsic

property of SMA to “remember” the high-temperature shape; however, additional training is required for SMA to “remember” the low-temperature shape.

In addition to the SME, SMAs have another unique property known as superelasticity, or pseudoelasticity, which occurs at temperatures above A_f and distinguishes SMA from other metallic alloys. Under external mechanical loading, detwinned martensite nucleates in the austenite matrix after reaching the critical resolved shear stress. Martensite continuously propagates with increasing external strain in a Lüders-type manner under a nearly constant stress (see the upper stress plateau in Fig. 3). The nucleation of martensite is so concentrated that a region nearly fully transformed to martensite is the first to form in the SMA. There is a clear edge between the martensite and austenite, with a shear angle of about 55° to the loading axis [15]. The martensite band propagates by new nucleation of martensite ahead of the martensite–austenite edge [16]. The inhomogeneous martensitic transformation is similar to the stress-induced slip band in low-carbon steels. This transformation behavior was first

reported by Guillaume Piobert and W. Lüders in 1864, and the martensitic band is then referred to as a Lüders-like deformation band (LBD).

The stress-induced martensite is unstable at the testing temperature (*higher than Af*) and will transform back to austenite when the external stress is unloaded. The superelasticity of SMA is associated with the reverse martensitic transformation and recovery of the macroscopic deformation strain. Upon unloading to the lower stress plateau (see Fig. 3), the martensite plates inside the LDB gradually transform back to austenite. At the macroscopic scale, the reverse transformation is characterized by shrinkage of the LDB.

Cross-References

- [Nanomechanical Properties of Nanostructures](#)
- [Plasticity Theory at Small Scales](#)
- [Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures](#)
- [Surface Tension Effects of Nanostructures](#)

References

1. Otsuka, K., Wayman, C.M.: Mechanism of shape memory effect and superelasticity. In: Otsuka, K., Wayman, C.M. (eds.) *Shape Memory Materials*, pp. 27–48. Cambridge University Press, Cambridge (1998)
2. Saburi, T., Nenno, S.: The shape memory effect and related phenomena. In: Proceedings of the International Conference on Solid-Solid Phase Transitions, pp. 1455–1479. Pittsburg (1981)
3. Otsuka, K., Ren, X.: Physical metallurgy of Ti–Ni-based shape memory alloys. *Prog. Mater. Sci.* **50**, 511–678 (2005)
4. Tanaka, Y., Himuro, Y., Kainuma, R., Sutou, Y., Omori, T., Ishida, K.: Ferrous polycrystalline shape-memory alloy showing huge superelasticity. *Science* **327**, 1488–1490 (2010)
5. Yue, Y.H., Liu, P., Zhang, Z., Han, X.D., Ma, E.: Approaching the theoretical elastic limit in Cu nanowires. *Nano Lett.* **11**, 3151 (2011)
6. Wong, E.W., Sheehan, P.E., Lieber, C.M.: Nanobeam mechanics: elasticity, strength and toughness of nanorods and nanotubes. *Science* **277**, 1971 (1997)
7. Park, H., Gall, S.K., Zimmerman, J.A.: Shape memory and pseudoelasticity in metal nanowires. *Phys. Rev. Lett.* **95**, 255504 (2005)
8. Humbeeck, J.V.: Non-medical applications of shape memory alloys. *Mater. Sci. Eng. A* **273–275**, 134–148 (1999)

9. Duerig, T., Pelton, A., Stöckel, D.: An overview of nitinol medical applications. *Mater. Sci. Eng. A* **273–275**, 149–160 (1999)
10. Wollants, P., Roos, J.R., Delaey, L.: Thermally- and stress-induced thermoelastic martensitic transformations in the reference frame of equilibrium thermodynamics. *Prog. Mater. Sci.* **37**, 227–288 (1993)
11. Liu, Y., McCormick, P.G.: Thermodynamic analysis of the martensitic transformation in NiTi – I. Effect of heat treatment on transformation behaviour. *Acta Metall. Mater.* **42**, 2401–2406 (1994)
12. Gall, K., Sehitoglu, H.: The role of texture in tension–compression asymmetry in polycrystalline NiTi. *Int. J. Plast.* **15**, 69–92 (1999)
13. Matsumoto, O., Miyazaki, S., Otsuka, K., Tamura, H.: Crystallography of martensitic transformation in Ti–Ni single crystals. *Acta Metall.* **35**, 2137–2144 (1987)
14. Wollants, P., Bonte, M.D., Roos, J.R.: A thermodynamic analysis of the stress-induced martensitic transformation in a single crystal. *Z. Metallkd.* **70**, 113–117 (1979)
15. Shaw, J.A.: Thermomechanical simulations of localized thermo-mechanical behavior in a NiTi shape memory alloy. *Int. J. Plast.* **16**, 541–562 (2000)
16. Mao, S.C., Luo, J.F., Zhang, Z., Wu, M.H., Liu, Y., Han, X.D.: EBSD studies of the stress-induced B2–B1' martensitic transformation in NiTi tubes under uniaxial tension and compression. *Acta Mater.* **58**, 3357–3366 (2010)

Superhydrophobicity

- [Lotus Effect](#)

Superoleophobicity of Fish Scales

Lei Jiang¹ and Ling Lin²

¹Center of Molecular Sciences, Institute of Chemistry Chinese Academy of Sciences, Beijing, People's Republic of China

²Beijing National Laboratory for Molecular Sciences (BNLMS), Key Laboratory of Organic Solids, Institute of Chemistry Chinese Academy of Sciences, Beijing, People's Republic of China

Synonyms

- [Oil-repellency of fish scales](#)

Definition

Superoleophobicity of fish scales is a wetting phenomenon in which fish scales show oil repellency in water, with a static oil contact angle (CA) higher than 150°. Generally, underwater superoleophobicity is defined as a static oil CA higher than 150° on a solid surface in an oil/water/solid three-phase system.

Chemical and Physical Principles

Wettability is a fundamental property of solid surfaces, which not only affects the behavior of the creatures in nature but also plays an important role in all aspects of our life. A direct expression of wetting behavior is a static CA of a liquid droplet sitting on a solid surface when the surface tensions at multiphase interfaces reach thermodynamic equilibrium.

For an ideal flat surface in a liquid/gas/solid system, the CA (θ) is given by the Young's equation [1]:

$$\cos \theta = \frac{\lambda_{sg} + \lambda_{sl}}{\lambda_{lg}} \quad (1)$$

where λ_{sg} is the solid/gas interface tension, λ_{sl} is the solid/liquid interface tension, and λ_{lg} is the liquid/gas interface tension.

For a rough surface in the air atmosphere, the situation is more complex. The surface topographic structure has a great influence on the wettability. Two distinct models, Wenzel's model and Cassie-Baxter's model, are commonly used to explain the effect of roughness on the apparent CAs of liquid drops. As described by Wenzel's model [2], liquid completely penetrates into the rough structures, leading to an increase in the contact area of the solid/liquid interface. Therefore, the apparent CA (θ_w) can be described as

$$\cos \theta_w = r \cos \theta \quad (2)$$

where the surface roughness factor (r) is defined as the ratio of the actual contact area of solid/liquid interface to the projected contact area

($r \geq 1$) and θ is the intrinsic CA on the flat surface. As described by Cassie-Baxter's model [3], liquid suspends on the rough surface rather than completely penetrates. The rough surface therefore can be considered as a solid/gas composite surface, and the apparent CA (θ_c) can be described as

$$\cos \theta_c = f \cos \theta - (1 - f) \quad (3)$$

where f is the area fraction of the solid/liquid interface and $(1 - f)$ is that of the liquid/gas interface. Based on both theories, scientists can better understand different wetting phenomena in nature, which would help the design of artificial materials with functional surfaces [4–6].

Key Research Findings

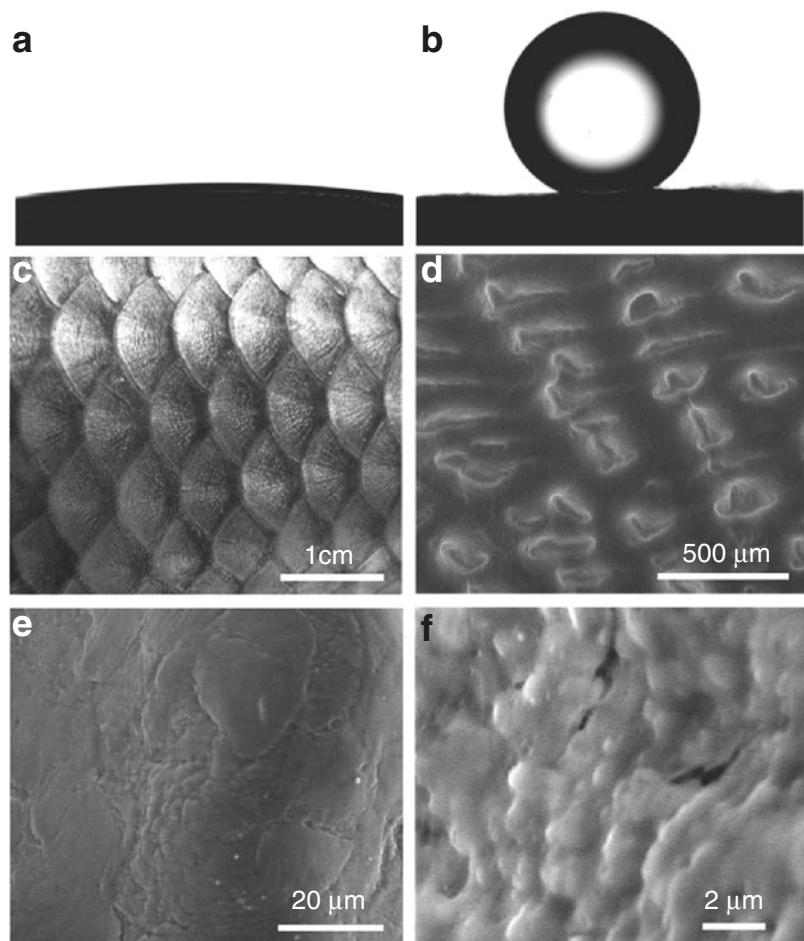
Oil-Wetting Behaviors on Fish Scales with Micro-/Nanostructures

Nature abounds with mysterious living organisms which have special surface properties. A famous example is the lotus leaf, a typical superhydrophobic surface in nature. When a water droplet falls on the lotus leaf, it keeps its bead shape and rolls off the surface immediately taking away the adherent dirt particles. This so-called lotus effect has been understood as a result of the complementary roles of low-surface free energy and micro-/nanostructures on the leaf surface [8]. Inspired by this phenomenon, tremendous artificial superhydrophobic surfaces have been fabricated and have facilitated practical applications in broad fields [9–11].

In a similar manner to the lotus effect in the air atmosphere, fish can resist oil pollution in water, possessing an antifouling mechanism. This fascinating phenomenon shows great potential for many applications in an oil/water/solid system, such as marine antifouling, prevention of oil spills, microfluidic technology, and bioadhesion [12, 13]. Liu et al. first reported the oil-wetting behavior on fish scales in the water environment [7]. It is common knowledge that fish scales are covered by a thin layer of mucus that leads to their hydrophilic nature. Liu et al. revealed that the

Superoleophobicity of Fish Scales, Fig. 1 (a)

Fish scales show superoleophilicity in air (1,2-dichloroethane (*DCE*) as a detecting oil, density 1.245 g cm^{-3} , surface tension at 25°C 31.86 mN m^{-1}). (b) Fish scales become superoleophobic once they are immersed in water (droplet of *DCE* as in (a)). (c–f) SEM images of fish scales disclose the surface micro-/nanostructures with the increase of magnification (Reproduced with permission. Copyright Wiley-VCH Verlag GmbH & Co. KGaA (2009))



hydrophilic surface of fish scales showed superoleophilicity in air (Fig. 1a) [14]. However, the surface of fish scales turned to be superoleophobic once it was immersed in water, with an oil CA larger than 150° (Fig. 1b). To figure out the reason for oil-wetting reversion, they firstly observed surface structures on fish scales in detail. Figure 1c–f display typical images of fish scales (crucian carp, *Carassius carassius*) using scanning electronic microscopy (SEM). The fan-shaped fish scales with diameters of 4–5 mm are densely arranged (Fig. 1c). The magnified images disclose that there are oriented micropapillae on each fish scale. Each micropapilla is in the length of 100–300 μm and in width of 30–40 μm (Fig. 1d). In high-magnification SEM images (Fig. 1e, f), nanoscale roughness is clearly observed on the surface of

micropapillae. It was suggested that these hierarchical structures could trap water and form a composite interface on fish scales to resist oil, which might play an important role in the oil-wetting reversion.

Mechanism of Wetting Behaviors on Fish Scales in Oil/Water/Solid System

To better understand the reversion of oil-wetting behavior on fish scales, Liu et al. chose three kinds of designed surfaces of silicon wafers as models: smooth surface, microstructured surface, and micro-/nanostructured surface [7]. Table 1 lists the contact angles of oil (*DCE*) or water on these three solid surfaces in the air or water environment, respectively. In air, the smooth surface is hydrophilic with a CA of $52.5 \pm 1.4^\circ$, while microstructured and micro-/nanostructured

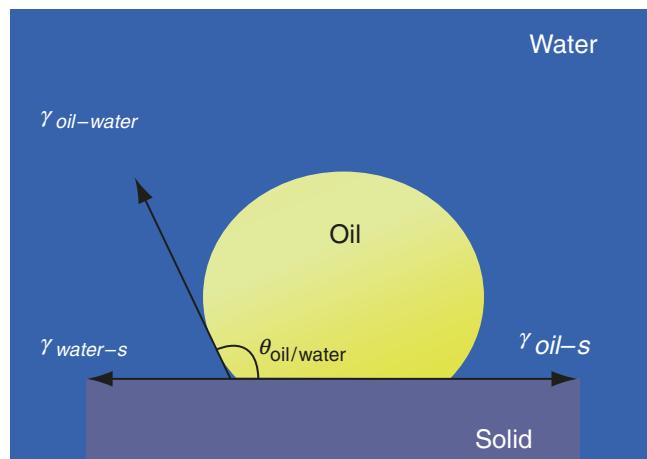
Superoleophobicity of Fish Scales, Table 1 Contact angles of oil (1,2-dichloroethane, DCE) or water on three solid surfaces with different surface structures in the air or water environment, respectively

Surfaces systems	Smooth silicon	Microstructured silicon	Micro-/nanostructured silicon
C ₂ H ₄ Cl ₂ droplets (in air)	<5°	<5°	<5°
Water droplets (in air)	52.5 ± 1.4°	<5°	<5°
C ₂ H ₄ Cl ₂ droplets (in water)	134.8 ± 1.6°	151.5 ± 1.8°	174.8 ± 2.3°

Source: Reproduced with permission. Copyright Wiley-VCH Verlag GmbH & Co. KGaA (2009)

Superoleophobicity of Fish Scales

Fig. 2 Schematic illustration of Young's equation in an oil/water/solid three-phase system, where a liquid droplet (*oil*) sits on a smooth surface (*solid*) in another liquid (*water*) phase



surfaces are both superhydrophilic (CAs < 5°). For the oil-wetting behaviors, all surfaces are superoleophilic, with the oil CAs smaller than 5°. However, in the water environment, all three kinds of surfaces become oleophobic or even superoleophobic. For the smooth surface, the oil CA is 134.8 ± 1.6°; for the microstructured surface, the oil CA is 151.5 ± 1.8°; and for the micro-/nanostructured surface, the oil CA is 174.8 ± 2.3°. The reversions of oil-wetting behaviors on these three surfaces are similar to that on fish scales. Comparing the different environments of two systems, the oil-wetting reversion is likely caused by the surrounding water. With these model silicon surfaces, the mechanism can be further analyzed through classical theories.

Although Young's equation was originally applied in a liquid/gas/solid three-phase system, it can be extended to an oil/water/solid system, in which an oil droplet sits on a solid surface in a water environment (Fig. 2). In this case, Young's equation is expressed as follows:

$$\cos \theta_{\text{oil/water}} = \frac{\gamma_{\text{water-s}} - \gamma_{\text{oil-s}}}{\gamma_{\text{oil-water}}} \quad (4)$$

where $\theta_{\text{oil/water}}$ is an oil CA on solid surface in an oil/water/solid three-phase system, $\gamma_{\text{water-s}}$ is the water/solid interface tension, $\gamma_{\text{oil-s}}$ is the oil/solid interface tension, and $\gamma_{\text{oil-water}}$ is the oil/water interface tension. Considering the liquid/gas/solid system, $\gamma_{\text{oil-s}}$ and $\gamma_{\text{water-s}}$ can be described as

$$\gamma_{\text{s-g}} = \gamma_{\text{oil-s}} + \gamma_{\text{oil-g}} \cos \theta_{\text{oil}} \quad (5)$$

$$\gamma_{\text{s-g}} = \gamma_{\text{water-s}} + \gamma_{\text{water-g}} \cos \theta_{\text{water}} \quad (6)$$

where θ_{oil} , θ_{water} is the oil or water CA in the air atmosphere, respectively, $\gamma_{\text{oil-g}}$ is the oil/gas interface tension, and $\gamma_{\text{water-g}}$ is the water/gas interface tension.

So Eq. 4 can also be expressed as follows:

$$\cos \theta_{\text{oil/water}} = \frac{\gamma_{\text{oil-g}} \cos \theta_{\text{oil}} - \gamma_{\text{water-g}} \cos \theta_{\text{water}}}{\gamma_{\text{oil-water}}} \quad (7)$$

Through Eq. 7, it can be easily understood why an oleophilic surface in air becomes oleophobic in water. Taking DCE as an example, the surface tension ($\gamma_{\text{oil-g}}$) of DCE is 24.1 mN m^{-1} , the water surface tension ($\gamma_{\text{water-g}}$) is 73 mN m^{-1} , and the interfacial tension ($\gamma_{\text{oil-water}}$) of DCE/water is 28.1 mN m^{-1} [14]. As mentioned in the above experimental results, in air, the water CA (θ_{water}) on the smooth silicon surface is $52.5 \pm 1.4^\circ$ and the oil CA (θ_{oil}) nearly 0° . Therefore, it can be calculated that $\cos \theta_{\text{oil/water}} = -0.72$ and $\theta_{\text{oil/water}} \approx 136^\circ$. This result implies that the reversion of oil-wetting behavior certainly happens when different three-phase systems are involved, which is consistent with the experimental oil CA ($134.8 \pm 1.6^\circ$) in water.

Moreover, when it comes to micro-/nanostructured surfaces, the oil CA is up to $174.8 \pm 2.3^\circ$ in the oil/water/solid system. This phenomenon is likely caused by the new forming interface on the micro-/nanostructured surface. According to the Cassie-Baxter's model, micro-/nanostructures could trap air forming a composite solid/gas surface in a liquid/gas/solid three-phase system. As to oil/water/solid

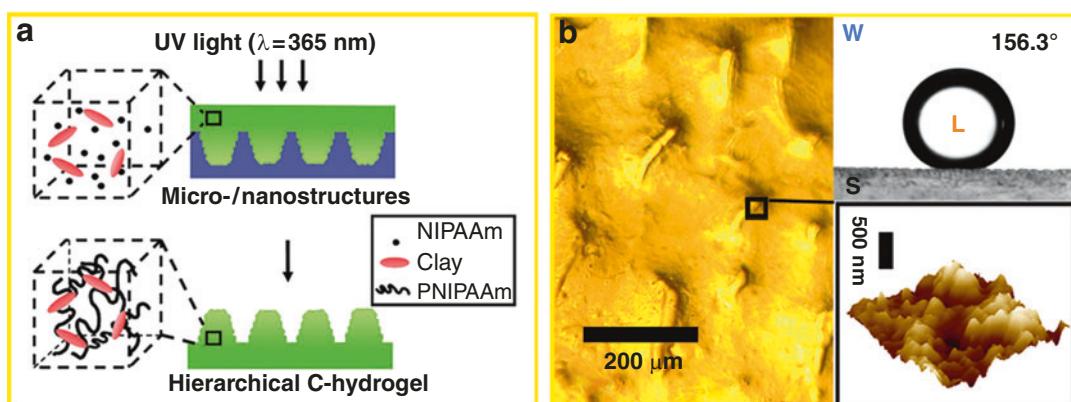
three-phase system, hierarchical structures can also trap abundant water forming a composite water/solid interface; therefore, the apparent oil CA ($\cos \theta'_{\text{oil/water}}$) in the water environment can be extended as

$$\cos \theta'_{\text{oil/water}} = f' \cos \theta_{\text{oil/water}} - (1 - f') \quad (8)$$

where f' is the area fraction of oil/solid interface, $(1 - f')$ is that of the oil/water interface, and $\theta_{\text{oil/water}}$ is the intrinsic CA of oil on a flat surface in oil/water/solid system. Once the composite surface forms, the oil droplet can rarely touch the solid surface, leading to the large increase of apparent oil CA. This effect endows the micro-/nanostructured silicon surface with superoleophobicity and as well contributes to the oil resistance of micro-/nanostructured fish scales.

Biomimetic Hydrogels for Robust Underwater Superoleophobicity

By understanding the fish oil-repellency nature, Lin et al. began to design a bionic artificial surface with robust underwater superoleophobicity [15]. From a practical perspective, the robustness of superoleophobic surface is crucial for underwater applications. Inspired by fish scales,

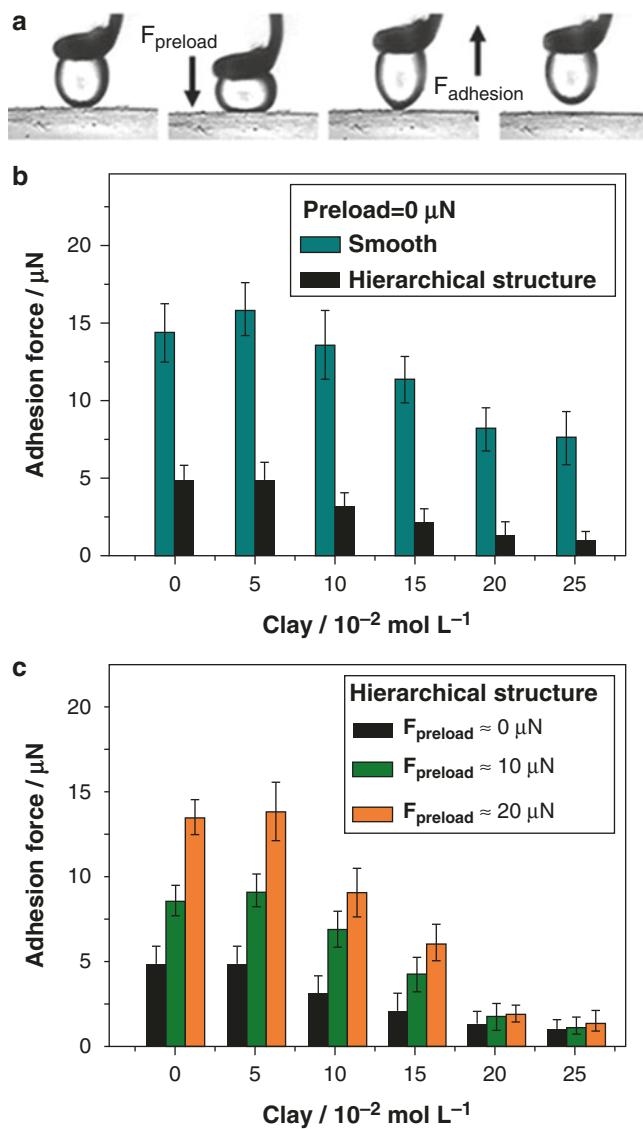


Superoleophobicity of Fish Scales, Fig. 3 Biomimetic design of PNIPAAm-nanoclay hydrogels (C-hydrogels) with hierarchical surface structures. (a) Schematic illustration of the fabricating process. (b) Optical and AFM images indicate micro-/nanostructured surface of fish

scale replica which shows underwater superoleophilicity. S = hydrogel; L = 1, 2-dichloroethane (DCE); W = water (Reproduced with permission. Copyright Wiley-VCH Verlag GmbH & Co. KGaA (2010))

Superoleophobicity of Fish Scales

Fig. 4 Robustness of superoleophobicity on C-hydrogel surfaces measured by dynamic underwater oil-adhesion measurements (oil: DCE). (a) A measurement process. (b) Oil-adhesion forces on hierarchical surfaces are much lower than those on smooth surface (preload = 0). (c) As the preload increases, the oil-adhesion force increases only on the low clay content C-hydrogel, while C-hydrogel with high clay content retains ultralow oil adhesion (Reproduced with permission. Copyright Wiley-VCH Verlag GmbH & Co. KGaA (2010))



hydrophilic hydrogel, biophysically similar to mucus, was chosen to construct a fish scale-like surface with micro-/nanostructures. Meanwhile, hydrophilic clay (synthetic hectorite), a rigid nanolayered structure, was selected as a composite component to enhance mechanical strength. In the experiment, the mixture of N-isopropylacrylamide (NIPAAm) and nanoclay was poured onto a polydimethylsiloxane (PDMS) template which was firstly molded from dried fish scales (grass carp, *Ctenopharyngodon idella*) (Fig. 3a). Then, a photo-initiated in situ

radical polymerization ($\gamma = 365 \text{ nm}$, 40 min) was taken to fabricate hybrid PNIPAAm-nanoclay hydrogels (C-hydrogels) with hierarchical surface structures. Figure 3b shows that micro-/nanostructures are well molded on the surface of the fish scale replica. Furthermore, note that the surface exhibits superoleophobicity with a static oil (CA) of $156.3 \pm 1.4^\circ$ in an oil/water/solid three-phase system (DCE as the detecting oil). This result showed that hybrid hydrogels with hierarchical surface structures successfully mimicked the oil-repellent fish scales.

To further investigate the robustness of superoleophobicity in a complex water environment, underwater oil-adhesion experiments were dynamically measured by a high sensitivity microelectromechanical balance system [13]. In each process, an oil droplet (DCE) was squeezed against the surface with a constant preload and then released, during which the adhesion force between oil and surface was recorded (Fig. 4a). It was found that, in the case of no preload on surfaces, the adhesion forces between oil and hierarchical surfaces were much lower than those between oil and smooth surfaces (Fig. 4b). As the preload increased, the oil-adhesion behaviors on hierarchical surfaces changed. The adhesion force increased only on the surface of low clay content C-hydrogel, while the C-hydrogel surface with high clay content retained excellent low oil adhesion (Fig. 4c). These results were attributed to the synergistic effects of rigid nanoclays and flexible macromolecules. It is known that the PNIPAAm possesses hydrophilic macromolecular chains which can trap water to prevent oil adhesion. But this effect is weakened by the fragility of hierarchical structures on the PNIPAAm surface. On the other hand, hybrid hydrogels with rigid nanoclays can enhance mechanical strength of surface micro-/nanostructures and, thus, keep the stability of trapped water on the surface. As a result, hybrid hydrogels with high clay content achieve robust underwater superoleophobicity. This study brings a new concept to the design and fabrication of underwater antifouling materials.

Future Directions for Research

The research on superoleophobic fish scales may open a new branch of the wettability field and make a strong impact on underwater applications. The studies in oil/water/solid three-phase systems have just begun, and many challenges remain on their way to development. First of all, the wetting theories in oil/water/solid system need to be further established. It is based on the tremendous wettability data of different materials with specific chemical components and surface structures.

Secondly, inspired by fish scale's effect, underwater superoleophobic materials with environment-friendly, durable properties need be further explored, which show great potential for underwater antifouling application. Finally, in oil/water/solid system, intelligent responsive materials with switchable oil-wetting behavior will attract great attention due to the promising applications in artificial muscles, actuators, and sensors. In the future, learning from nature will be a primary principle to design biomimetic or bioinspired functional materials with special wettability.

Cross-References

- [Biomimetics of Marine Adhesives](#)
- [Shark Skin Effect](#)

References

- Young, T.: An essay on the cohesion of fluids. *Philos. Trans. R. Soc. Lond. A* **95**, 65–87 (1805)
- Wenzel, R.N.: Resistance of solid surfaces to wetting by water. *Ind. Eng. Chem.* **28**, 988–994 (1936)
- Cassie, A.B..D., Baxter, S.: Wettability of porous surfaces. *Trans. Faraday Soc.* **40**, 0546–0550 (1944)
- Öner, D., McCarthy, T.J.: Ultrahydrophobic surfaces. Effects of topography length scales on wettability. *Langmuir* **16**(20), 7777–7782 (2000)
- Extrand, C.W.: Model for contact angles and hysteresis on rough and ultraphobic surfaces. *Langmuir* **18**(21), 7991–7999 (2002)
- Quéré, D.: Wetting and roughness. *Annu. Rev. Mater. Res.* **38**, 71–99 (2008)
- Liu, M.J., Wang, S.T., Wei, Z.X., Song, Y.L., Jiang, L.: Bioinspired design of a superoleophobic and low adhesive water/solid interface. *Adv. Mater.* **21**(6), 665–669 (2009)
- Barthlott, W., Neinhuis, C.: Purity of the sacred lotus, or escape from contamination in biological surfaces. *Planta* **202**(1), 1–8 (1997)
- Blossey, R.: Self-cleaning surfaces – virtual realities. *Nat. Mater.* **2**(5), 301–306 (2003)
- Gennes, P.-G., Brochard-Wyart, F., Quéré, D.: Capillarity and Wetting Phenomena: Drops, Bubbles, Pearls, Waves. Springer, New York (2004)
- Yao, X., Song, Y.L., Jiang, L.: Recent developments in bio-inspired special wettability. *Adv. Mater.* **23**, 719–734 (2011)
- Nosonovsky, M., Bhushan, B.: Multiscale effects and capillary interactions in functional biomimetic

- surfaces for energy conversion and green engineering. *Philos. Trans. R. Soc. Lond. A* **367**(1893), 1511–1539 (2009)
- 13. Liu, M.J., Zheng, Y.M., Zhai, J., Jiang, L.: Bioinspired super-antiwetting interfaces with special liquid–solid adhesion. *Acc. Chem. Res.* **43**(3), 368–377 (2010)
 - 14. Lide, D.R.: CRC Handbook of Chemistry and Physics, 84th edn. CRC Press, Boca Raton (2003–2004)
 - 15. Lin, L., Liu, M.J., Chen, L., Chen, P.P., Ma, J., Han, D., Jiang, L.: Bio-inspired hierarchical macromolecule-nanoclay hydrogels for robust underwater superoleophobicity. *Adv. Mater.* **22**(43), 4826–4830 (2010)

Superparamagnetism

- Magnetic Nanoparticles for Biomedical Applications

Support Loss

- Anchor Loss in MEMS/NEMS

Surface Dissipations in NEMS/MEMS

Jinling Yang
Institute of Semiconductors, Chinese Academy of Sciences, Beijing, People's Republic of China
State Key Laboratory of Transducer Technology, Shanghai, People's Republic of China

Synonyms

Surface loss in micromechanical/nanomechanical resonators; Surface loss in NEMS/MEMS; Mechanical energy dissipation

Definition

Surface dissipation is the mechanical energy loss caused by surface defects, such as dangling bonds,

absorbates, and crystal termination defects. It becomes dominant as the dimensions of nanoelectromechanical systems (NEMS)/microelectromechanical systems (MEMS) resonators are reduced and the surface-to-volume ratio grows.

Overview

Nanoelectromechanical systems (NEMS)/microelectromechanical systems (MEMS) are systems integrating nanometer/micrometer-scale mechanical and electrical components. NEMS/MEMS resonators play an important role in viable commercial technologies and are becoming more and more prevalent in research applications; for example, micromechanical resonators are excellent transducers for force or mass detection [1, 2]. Advances in nanofabrication technology have enabled extreme miniaturization of resonant sensors. As tools for basic research, NEMS resonators have demonstrated extraordinary sensitivity to external forces, allowing the detection of electron spin flips [3], single molecules [4], and other fundamental phenomena [5, 6].

The fundamental characteristics of a mechanical resonator are determined by the resonance frequency and quality factor Q (inverse dissipation). Better device response is obtained by high frequency and high Q . Scaling down the dimensions is necessary for achieving high resonance frequencies. However, the expected gains in resonance frequencies from size reduction beyond the submicron scale have often been offset by degradation in quality factor. With the ever-increasing technical ability, it has become possible to miniaturize the resonator dimension to several tens of nanometer. Correspondingly, many different dissipation sources have been identified in the NEMS/MEMS device.

Dissipation effects appear in all mechanical systems and help define fundamental dynamical behavior. Understanding this issue is important, not only for improving the resonator mechanical properties but also for establishing their performance limits. Low dissipation is generally desirable, which makes a device more efficient and more sensitive and less susceptible to wear and mechanical noise [7]. Dissipation studies of

miniaturized resonators have been conducted for decades. The factors that degrade the Q value can be categorized as intrinsic or extrinsic. Extrinsic damping is attributed to interactions of the NEMS/MEMS with its surrounding environment, i.e., the gas molecules surrounding it, or the substrate on which it lies. Intrinsic damping results from flaws or defects in the NEMS/MEMS. There are essentially four major loss mechanisms for NEMS/MEMS. Extrinsic losses include loss in airflow and radiation of elastic wave at the support (support/clamping loss). Intrinsic losses are thermoelastic loss and surface loss [8].

Dissipation is a measure of energy lost per oscillation in the resonator. The Q -factor can be described either as the fullwidth at half-maximum of the measured resonance peak or the rate at which the NEMS/MEMS loses energy per vibrational period. The resonator Q is defined as $Q = 2\pi W_0 / \Delta W$, where W_0 is the stored vibrational energy and ΔW is the total energy lost per cycle of vibration. ΔW can be written as $\Delta W = \sum_i \Delta W_i$, where ΔW_i represents the energy lost due to the various dissipation mechanisms. Every one of the abovementioned loss mechanism has an associated Q factor; the overall quality factor Q_{tot} can be found from [9]

$$\frac{1}{Q_{tot}} = \sum_i \frac{1}{Q_i}, \quad (1)$$

It is obvious that Q_{tot} cannot exceed the smallest Q_i .

As the resonators become thinner (or narrower), the surface-to-volume ratio grows. The surface contains a large amount of defects due to the lattice termination and surface impurities, which contribute significantly to damping in nanoresonators. Eventually, the surface properties play a dominant role in the dissipation over the bulk behaviors.

Key Principle

Surface Dissipation Mechanism

Cantilever is a typical NEMS/MEMS resonator; extensive studies have been done to the

mechanical behavior of the cantilever [8, 10–12]. Hereafter, the NEMS/MEMS cantilever will be used to describe the general dissipation behavior of NEMS/MEMS resonator. When the cantilever thickness scales down, the surface-to-volume ratio increases, the surface loss becomes dominant. The surface loss is caused by absorbates or surface defects. The charge transfer between the surface and the absorbates, the Coulomb repulsion of the dipole moments associated with the adsorbate atoms, and the overlap of the wave function of adsorbate atom orbits at a close distance would modify the surface stress and lead to surface loss [13–15].

The surface dissipation is modeled by considering the complex Young's modulus $E_c = E + iE_d$, where E_c and E_d are the complex value and the dissipation part of Young's modulus, respectively [12]. For a rectangular cantilever vibrating sinusoidally, the stored energy can be written as [16]

$$W_0 = \frac{1}{6} whE \int_0^l \varepsilon_{max}^2(x) dx, \quad (2)$$

where l , w , and h are the beam length, width, and thickness, respectively, ε_{max} is the strain occurring on the top or bottom surface of beam during vibration. Considering the surface layer with thickness δ and complex modulus $E_{cs} = E_s + iE_{ds}$, the energy loss per cycle due to the surface layer is [12]

$$\Delta W_s = 2\pi\delta E_{ds} \left(w + \frac{h}{3} \right) \int_0^l \varepsilon_{max}^2(x) dx, \quad (3)$$

where E_{cs} , E_s , and E_{ds} are the complex, conventional, and the dissipation values of the Young's modulus of the surface layer, respectively. Thus the surface loss related Q is

$$Q = \frac{wh}{2\delta(3w+h)} \frac{E}{E_{ds}}, \quad (4)$$

In Equation 4, E_{ds} , as a property of the adsorbate layer and its defect, is closely related to the surface

stress and results in the surface loss. For a thin cantilever with $h \ll w$, the surface-loss related Q factor would be proportional to the thickness. Therefore, in the high surface-to-volume ratio structure, surface effect has become a crucial restriction for resonator miniaturization. To gain deep insight into this dissipation mechanism and its relationship with the dimensions of devices, the surface effect has been investigated via surface treatment in various conditions.

Key Research Findings

The cantilevers, with lengths of 5–120 μm , have been fabricated from (100)-oriented SIMOX (Separated by IMplanted OXygen) wafers with 60- (Fig. 1) and 170-nm-thick top Si layers, SOI (Silicon On Insulator) wafers with a 500-nm-thick top Si layer, and (110)-oriented SIMOX wafers with a 160-nm-thick top Si layer [10, 11]. Length to width ratio for all the cantilevers is about 10:1, as shown in Fig. 1. The cantilevers are actuated by piezo. The resonance of the cantilevers was measured by laser Doppler vibrometer. All measurements are performed at high vacuum ($<10^{-5}$ Torr) in order to avoid air damping.

Surface dissipation is proportional to the square of the peak strain integrated over the surface area of the beam. For the fundamental

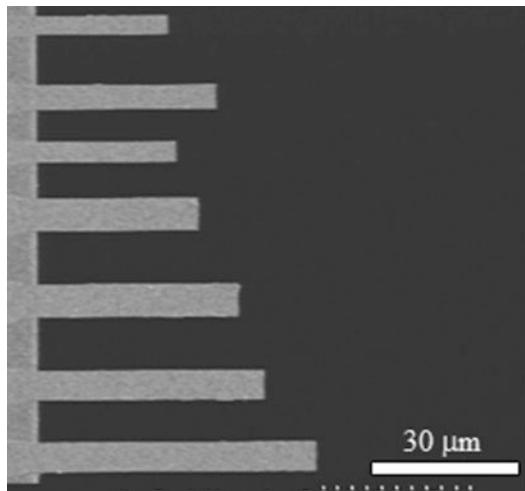
flexural mode of the cantilever, the Q factor would be independent of cantilever width, but proportional to the thickness because of the dominant surface dissipation [11]. Thus the surface effects can experimentally be quantified by changing the cantilever thickness. For very thin and long beams, surface dissipation is dominant until clamping losses increase sharply as the length is reduced. As shown in Fig. 2, for cantilevers with $l > 30 \mu\text{m}$, Q factor is almost proportional to the thickness, when the thickness of cantilevers is scaled up to 500 nm or down to 60 nm. But deviations from this relationship were found when $l < 30 \mu\text{m}$, suggesting that the support loss plays an important role in the short cantilevers [8]. The linear increase in the mechanical Q with the cantilever thickness was also observed in silicon–nitride cantilever of thicknesses 200 nm, 510 nm, 700 nm, and 1.2 μm [12].

Surface Treatments at High Temperature

Surface treatments such as high-temperature annealing, which removes adsorbates and releases strain, or chemical treatments have significant effects on the quality factor, and improvements over one order of magnitude have been reported [8, 10–12].

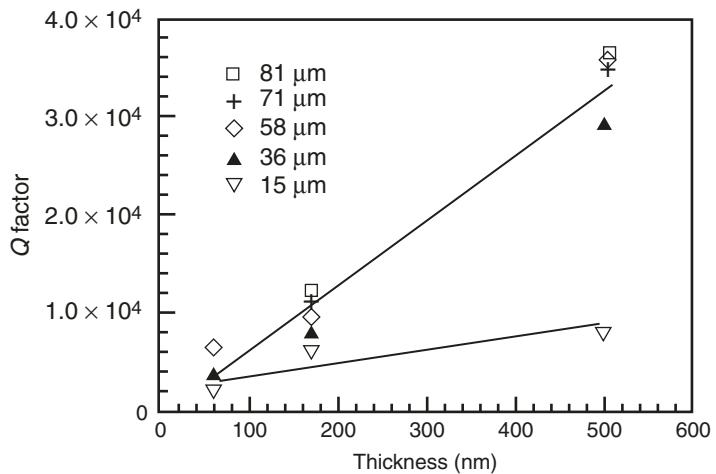
It is well known that the clean Si(100) surface usually displays a 2×1 structure [17]. On these bulk-terminated surfaces, each Si atom has one or two dangling bonds (dbs), which determines the initial reactivity of the surface and a final surface structure. In addition, hydrogen plays an important role in silicon surface chemistry [17, 18]. Thus the surface stress could be modified by hydrogen dose under various conditions. And hydrogen-terminated surface is chemically inert and is not easily oxidized in atmosphere.

In atmosphere, the Si cantilever surface is covered by native SiO_2 and adsorbates, which finally causes energy loss [10, 11]. The following treatment process was employed to modify the Si surface chemistry: firstly, annealing is done in the UHV chamber (1×10^{-10} mbar) to obtain a clean surface, that is, the cantilevers were outgassed at 600 °C for 30–60 min, then quickly



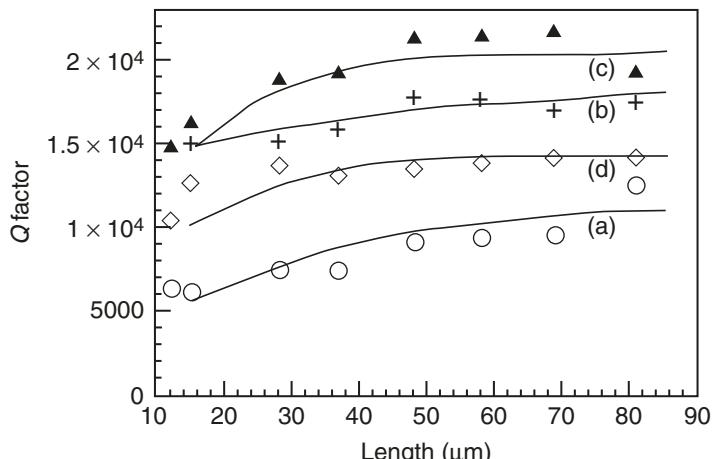
Surface Dissipations in NEMS/MEMS, Fig. 1 SEM image for cantilevers of 170 nm in thickness and of 10–100 μm in length

Surface Dissipations in NEMS/MEMS, Fig. 2 Q factors versus thickness for cantilevers with lengths of 15 μm , 36 μm , 58 μm , 71 μm , and 81 μm . All the measurements were done in the high vacuum system



Surface Dissipations in NEMS/MEMS

Fig. 3 Length dependence of Q factors for 170-nm-thick cantilevers before (a) and after (b) heating at 600 $^{\circ}\text{C}$ for 1 h and flashing at 1000 $^{\circ}\text{C}$ for three times, after exposure to atomic hydrogen above 300 $^{\circ}\text{C}$ for 30 s (c) and at room temperature for 2 min (d). The surface treatment was performed in UHV setup, and all the measurements were done in the high vacuum system



flashed at 900–1000 $^{\circ}\text{C}$ three times to remove SiO_2 ; secondly, exposure to atomic hydrogen is performed to passivate the clean surface, atomic hydrogen was produced by decomposition of molecule hydrogen on a 1500 $^{\circ}\text{C}$ tungsten filament. Exposure to atomic hydrogen with a pressure around 5×10^{-8} torr was accomplished at room temperature and above 300 $^{\circ}\text{C}$ for 30–120 s. After being cooled to room temperature in the UHV chamber, the cantilevers were transferred to a laser Doppler system for measurement. The time interval for this transfer was kept below 15 min to reduce the atmosphere effect on the surface.

As shown in Fig. 3, for 170-nm-thick Si(100) cantilever, after heating at 1000 $^{\circ}\text{C}$,

Q factors of all cantilevers clearly increase due to the removal of the SiO_2 layer and surface absorbates. After exposure to atomic hydrogen, the cantilever surface could further be modified, since hydrogen reacts with dbs on the Si dimer atoms, and the Si surface becomes inert, the growth of native SiO_2 and the absorption process during sample transfer slow down, thus, higher Q factors were achieved for the cantilevers (Fig. 3c). However, the extended exposure to atomic hydrogen for 2 min at room temperature results in serious etching of the cantilevers and deteriorates their Q factors very much (Fig. 3d). It is well established in [17, 18] that saturation exposure of a clean Si(100) 2×1 surface to atomic hydrogen at 300 $^{\circ}\text{C}$ results in the formation of a

monohydride surface, which preserves the Si dimer bonds and the associated 2×1 surface periodicity. Saturation exposure at lower temperature (room temperature) results in the 1×1 dihydride surface structure. The steric interaction between the neighboring dihydride units on the 1×1 surface produces strain which weakens the bonds within the units; as a result, these bonds are susceptible to further attack and etching by hydrogen atoms.

The surface modification process can be outlined as Fig. 4, annealing at 600°C in UHV could remove some absorbates from the cantilever surface, e.g., H_2O and some organics. Subsequent

flashing at 1000°C for a short time will further clean the surface by desorbing the organics and SiO_2 layer and modify the surface structure; some of the surface structures may convert to $\text{Si}(100)$ 2×1 with one dangling bond on each dimer atom. During transfer from the treatment chamber to the laser Doppler system, very thin native oxide and adsorbate layer could be formed on the clean surface. Exposure to atomic hydrogen at above 300°C is favorable for formation of $\text{Si}(100)$ 2×1 : H monohydride surface structure with high stability and less stress. The considerable increase of the Q factor after exposure to atomic hydrogen could be a consequence of both surface modifications by hydrogen and the passivation effect.

Surface dissipation is closely related to the crystallographic orientation with different surface defect densities. As shown in Fig. 5, the Q factors of the 160-nm-thick cantilevers made from (110)-oriented SOI wafer are comparable to the corresponding values of the 170-nm-thick cantilevers made from (100)-oriented SOI wafer. However, heating and exposure to atomic hydrogen increases the Q values of the 160-nm-thick cantilevers much less than the 170-nm-thick cantilevers. This distinction could be attributed to the different structures and defect density of two surfaces. After annealing at 1000°C , the clean $\text{Si}(110)$ surface could reconstruct into a 2×16 structure, and hydrogen exposure would form a bulk-like 2×16 : H surface structure [10]. The surface modification on $\text{Si}(110)$ by annealing and exposure to atomic hydrogen is far different from the $\text{Si}(100)$ structure.

Surface treatment has more influence on the Q factors of the thinner structures, as depicted in Fig. 6 for the relative change of Q values before and after exposure to atomic hydrogen ($Q_0 - Q_H$) / Q_0 [10].

For clarifying how large the surface dissipation is in the ultrathin cantilever, the atmosphere effect must be completely eliminated. This is important not only for the improvement of the Q factor but also for a stable operation of the cantilever dynamically under a certain environment. Therefore, an UHV system with a preparation chamber for surface treatment and a laser Doppler chamber for in situ mechanical properties measurement

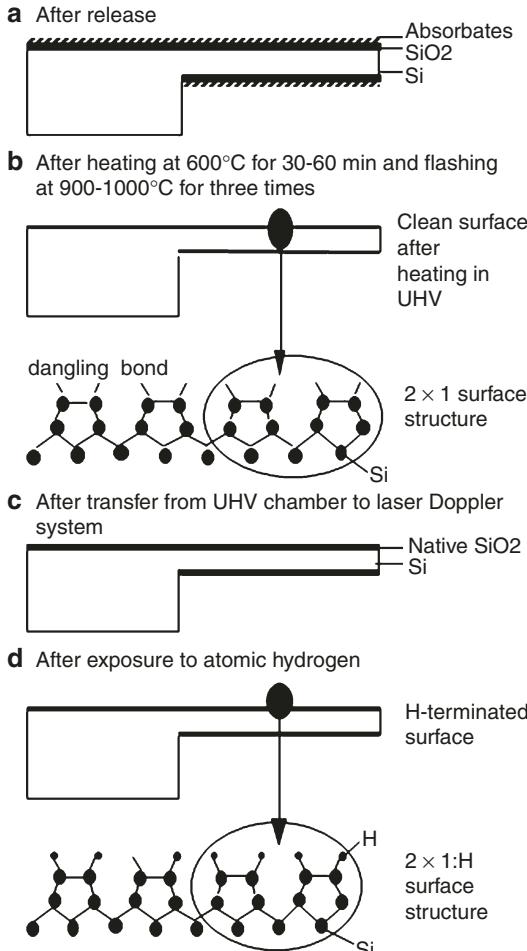
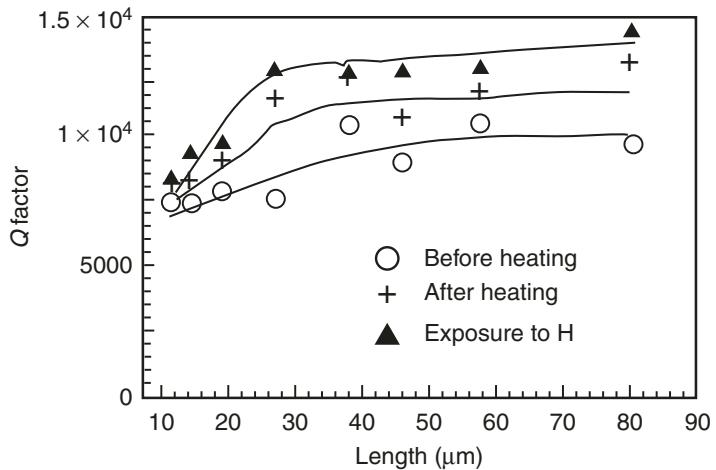


Fig. 4 Schematic drawing of surface treatment for $\text{Si}(100)$ resonators

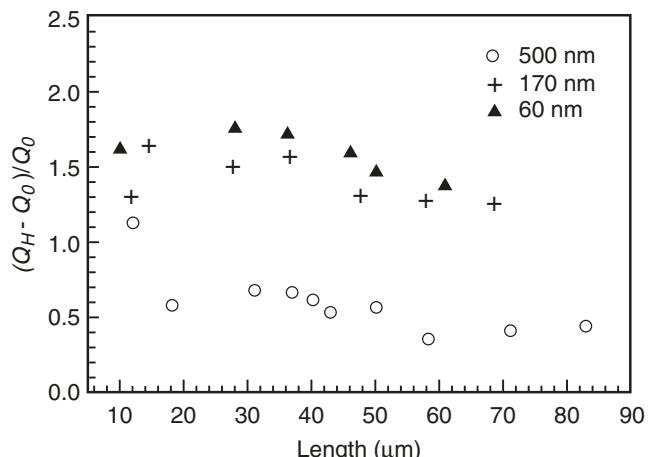
Surface Dissipations in NEMS/MEMS,

Fig. 5 Q -factors versus length for (110)-oriented cantilevers before and after heating, and after exposure to atomic hydrogen



Surface Dissipations in NEMS/MEMS,

Fig. 6 Length dependence of the relative change of Q factors before (Q_0) and after (Q_H) exposure to atomic hydrogen for cantilevers with thickness of 60 nm, 170 nm, and 500 nm. All the measurements were done in the high vacuum system



was employed [11], as shown in Fig. 7. The cantilever is optically actuated by a laser beam. The laser power dependence of the resonance frequency and the Q factor of a 48 μm long cantilever annealed at 1000 °C for 30 s was investigated and is shown in the inset of Fig. 8. A laser output power was limited to approximately 40 μW , which is enough to excite and maintain a stable resonant vibration for all cantilevers with different lengths and can ensure that the measurements were undisturbed by the “soft spring” effect.

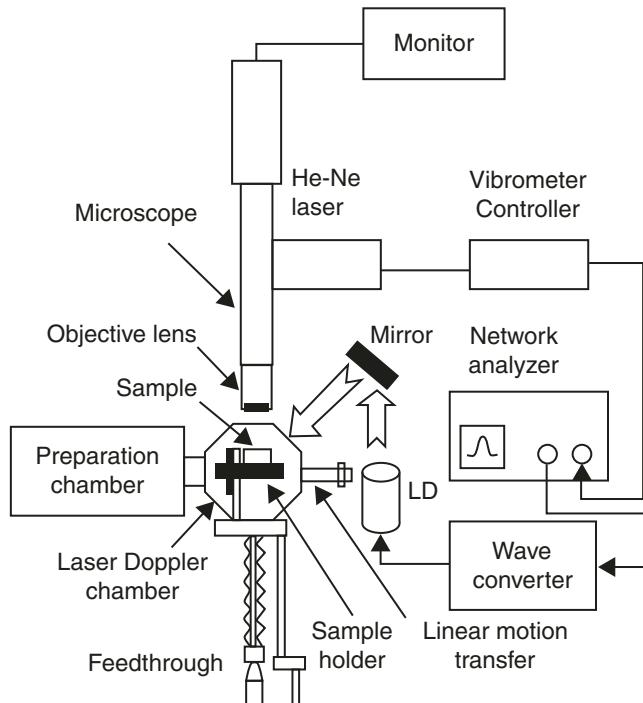
The good vacuum condition in UHV system enables the clean surfaces of the cantilevers to be preserved during measurement. After annealing at 600 °C or 1000 °C in the preparation chamber, the

sample was transferred to the laser Doppler chamber by a linear motion transfer. The mechanical resonance of the actuated cantilever is detected by the laser Doppler system and analyzed by a network analyzer, which also outputs a drive signal to modulate the pump frequency of the laser diode and thereby drives the cantilever into oscillation.

Figure 9 shows the length dependence of the Q factors for 170-nm-thick cantilevers under different treatments in the UHV chamber [11]. The freshly fabricated cantilevers of 30–90 μm long have the Q factors around 10^4 , with a little dependence on their length. The shorter cantilevers ($L < 30 \mu\text{m}$) have the lower Q factors due to the obvious support loss [8]. Annealing at 600 °C

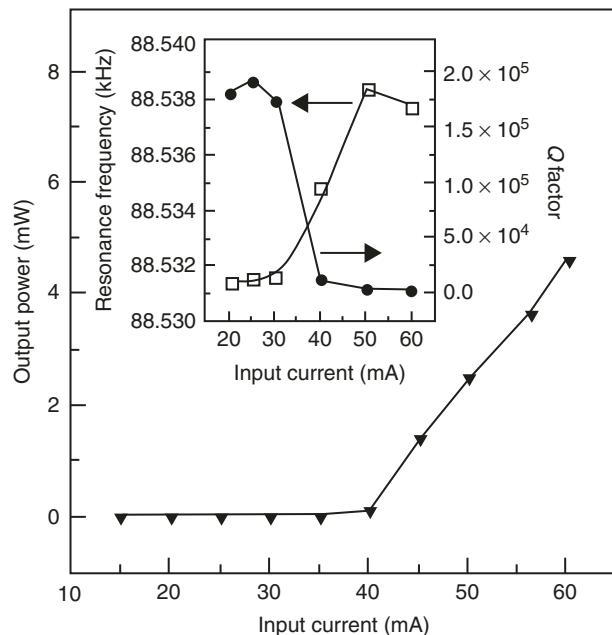
Surface Dissipations in NEMS/MEMS,

Fig. 7 Schematic diagram of UHV system consisting of a preparation chamber, a measurement chamber, and a laser Doppler measurement loop. The cantilevers are actuated optically



Surface Dissipations in NEMS/MEMS,

Fig. 8 Input current dependence of laser power before irradiation onto the UHV optic window. Inset is the measured resonance frequency and Q factor as a function of input current of LD for a cantilever of 48 μm long, 5 μm wide, and 170 nm thick. Soft spring effect induced sharp changes in the resonance frequency and Q factor can be clearly seen at 30–40 mA

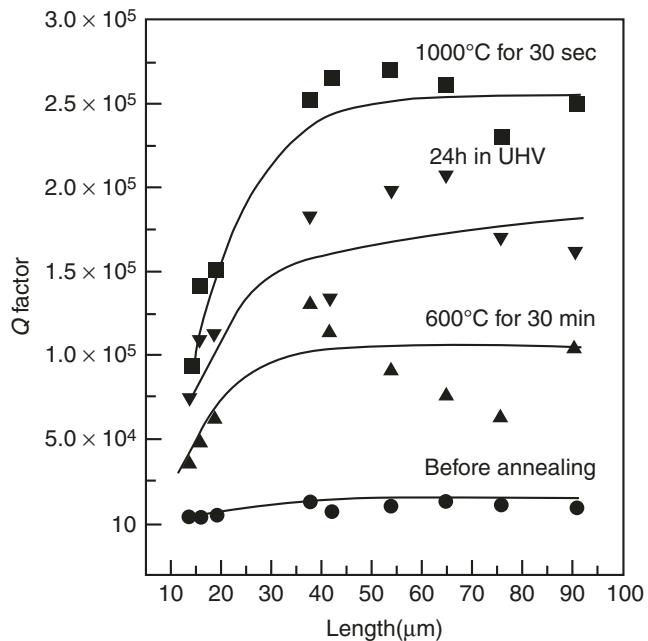


for 30 min in UHV increases the Q factor by one order of magnitude (into the 10^5 range). Subsequent annealing at 1000 °C for 30 s causes a further increase of the Q values by a factor of 2–3. The improvement of Q factor was found to

be associated with removing the absorbates and the deoxidization of the surface. Keeping the sample in UHV for about 24 h after annealing at 1000 °C obviously reduces the Q factors of all the cantilevers, which suggests that the change in

Surface Dissipations in NEMS/MEMS, Fig. 9

factors for a set of cantilevers fabricated on one chip, 170 nm in thickness and 13–90 μm in length. UHV annealing leads to significant increment on Q factors, up to $\sim 2.5 \times 10^5$ for cantilevers of 30–90 μm long annealed at 1000 °C for 30 s. All the annealing and measurements were done in the UHV system



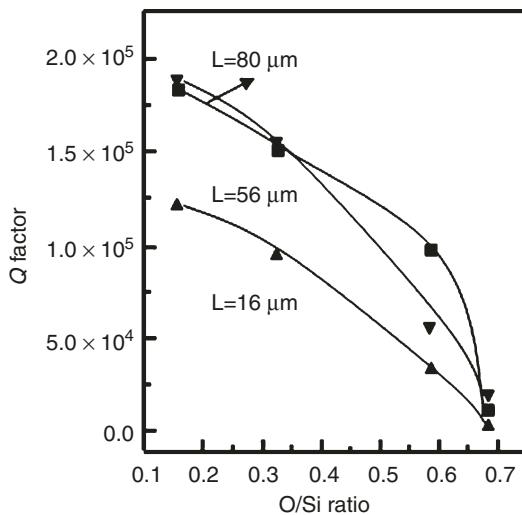
Q factor is caused by the surface modification rather than the annealing effect. These results indicate that the ultrathin cantilever is seriously subjected to the surface effect, that is, the surface treatment in UHV modifies the surface structure and thus the surface stress, which is related to the surface energy dissipation.

Moreover, in order to get further insight into the surface modification during annealing the samples, one large pattern of $400 \times 400 \mu\text{m}^2$ plus several microcantilevers of 170 nm thick and several tens of micrometers long were fabricated on one chip by the same process [11]. This patterned chip was annealed in UHV at 1000 °C for 15 s four times in a sequence. After each annealing, the surface of this square pattern was monitored by x-ray photoelectron spectroscopy (XPS) while the mechanical properties of the microcantilevers were examined. After annealing at 1000 °C for the first 15 s, no carbon peak could be detected, only photoelectron peaks of Si_{2S}, Si_{2P}, and O_{1S}, and Auger peak of O_{KLL} were observed. As shown in Fig. 10 for cantilevers of 16 μm , 56 μm , and 80 μm long, annealing at 1000 °C in UHV reduced oxygen concentration on the surface and simultaneously enhanced the Q values.

Chemical Passivation

Mechanical energy dissipation in micromechanical silicon structures is sensitive to the chemical state of the surface. By changing a single monolayer of molecules with less than 0.07 % of the total mass on the surface of a 5- μm -wide, 250-nm-thick Si(111) resonator, the resonator's quality factor can be improved by at least 70 %. When a single monolayer of hydrogen atoms is replaced by 13 Å of silicon oxide, corresponding to the oxidation of less than two silicon bilayers, which is standard for commercial silicon devices, mechanical energy dissipation increases significantly: at least 75 % of the energy in the oxide-terminated resonators is dissipated at the surface [19].

The mechanical properties of coated resonators are correlated with the chemical passivation properties of the monolayer. High-density monolayers (methyl groups) that resist the formation of electronic defects have the best performance. Resonators terminated with these passivation monolayers are also very stable in both vacuum and air due to their relatively low densities of electronic defects as well as the resistance to chemical attack [20].



Surface Dissipations in NEMS/MEMS,
Fig. 10 Monotonic dependence of Q factors on O/Si ratio for three cantilevers of 170 nm thick, 16 μ m, 56 μ m, and 80 μ m long, respectively, which are determined by in situ laser Doppler measurement and XPS after being annealed at 1000 $^{\circ}$ C for 15 s, 30 s, 45 s, and 60 s

Reducing Surface Dissipation

To improve device performance of NEMS/MEMS, methods have been developed to reduce surface dissipation. Here, we present two chemical methods that can have a profound influence on the dissipative nature of NEMS/MEMS structures. The first is a method where surface dissipation is reduced through annealing. The second is passivation of the surface with submonolayer.

Annealing at High Temperature

The atoms that lie at the surfaces of the NEMS/MEMS have fewer bonding neighbors than atoms in the bulk and are susceptible to surface contamination, which will reduce the quality factor of a NEMS/MEMS resonator. For silicon resonators, annealing in various conditions has proven to be effective for reducing the surface dissipation. Treatment at high temperature in UHV chamber was able to remove the oxide and absorbates on the resonator surface, modify the surface stress, and improve the Q values beyond 10^5 for

170-nm-thick cantilever. As expected, the thin structures are more sensitive to surface degradation and treatment than the thicker structures [10].

High-temperature annealing the 70 nm and 170-nm-thick single-crystal silicon cantilevers at 700 $^{\circ}$ C for 1 h in forming gas (Ar with 4.25 % H) and in a nitrogen atmosphere produced increases about a factor of two and three, respectively [12].

Passivating Resonator Surface

The clean Si(100) surface usually displays a 2×1 structure, each Si atom has one or two dangling bonds (dbs), which determines the initial reactivity of the surface and a final surface structure. The surface of the freshly fabricated resonators or the clean surface after annealing in UHV is easily subjected to surface contamination and energy loss in atmosphere, even in UHV the Q factors of the cantilevers degrades with time. Passivating surface atoms to make their bonding environments more bulk-like has resulted in higher Q factors [8, 19, 20].

Mechanical energy dissipation in a micromechanical silicon resonator is sensitive to submonolayer changes in surface chemistry. Hydrogen plays an important role in silicon surface chemistry and hydrogen-terminated surface is chemically inert. The effect of different alkenes and hydrogen surface termination on silicon resonators was studied [20]. The rate of mechanical energy dissipation and the density of electronically active defects on freshly prepared Si(111) resonators and functionalized surfaces follow the trend: silicon oxide $>>$ long-chain alkyl $>$ H $>$ CH₃. After an extended period of air exposure, H-terminated surfaces degrade significantly, leading to silicon oxide $>$ H $>$ long-chain alkyl $>$ CH₃. Si(111) resonators terminated by a single monolayer of methyl groups have significantly higher quality factors, and thus lower rates of mechanical energy dissipation, than those terminated with either long-chain alkyl monolayers (-C_nH_{2n+1}, n = 2–18) or hydrogen monolayers.

In summary, scaling down the size of the NEMS/MEMS resonators does not necessarily

mean the suppression of Q factor and that a higher Q factor and thus higher sensitivity are achievable if the proper surface treatment and passivation process are adopted for the NEMS/MEMS resonators.

Cross-References

- [Microcantilever Chemical and Biological Sensors](#)
- [MicroElectroMechanical Systems](#)
- [NEMS](#)
- [NEMS Resonant Mass Sensors](#)

References

1. Rugar, D., Zuger, O., Hoen, S., Yannoni, C.S., Vieth, H.M., Kendrick, R.D.: Force detection of nuclear magnetic resonance. *Science* **264**, 1560–1563 (1994)
2. Huang, X., Feng, X., Zorman, C., Mehregany, M., Roukes, M.: VHF, UHF and microwave frequency nanomechanical resonators. *New J. Phys.* **7**, 247 (2005)
3. Zolfagharkhani, G., Gaidarzhy, A., Degiovanni, P., Kettemann, S., Fulde, P., Mohanty, P.: Nanomechanical detection of itinerant electron spin flip. *Nat. Nanotechnol.* **3**, 720–723 (2008)
4. Naik, A., Hanay, M., Hiebert, W., Feng, X., Roukes, M.: Towards single-molecule nanomechanical mass spectrometry. *Nat. Nanotechnol.* **4**, 445–450 (2009)
5. Wu, G., Ji, H., Hansen, K., Thundat, T., Datar, R., Cote, R., Hagan, M., Chakraborty, A., Majumdar, A.: Origin of nanomechanical cantilever motion generated from biomolecular interactions. *Proc. Natl. Acad. Sci.* **98**, 1560–1564 (2001)
6. Montemagno, C., Bachand, G.: Constructing nanomechanical devices powered by biomolecular motors. *Nanotechnology* **10**, 225–231 (1999)
7. Cleland, A.: Themomechanical noise limits on parametric sensing with nanomechanical resonators. *New J. Phys.* **7**, 235 (2005)
8. Yang, J.L., Ono, T., Esashi, M.: Energy dissipation in submicrometer thick single-crystal silicon cantilevers. *IEEE J. Microelectromech. Syst.* **11**, 775–783 (2002)
9. Stemme, G.: Resonant silicon sensors. *J. Micromech. Microeng.* **1**, 113–125 (1991)
10. Yang, J.L., Ono, T., Esashi, M.: Investigating surface stress: surface loss in ultrathin single-crystal silicon cantilevers. *J. Vac. Sci. Technol. B* **19**, 551–556 (2001)
11. Yang, J.L., Ono, T., Esashi, M.: Surface effects and high quality factors in ultrathin single-crystal silicon cantilevers. *Appl. Phys. Lett.* **77**, 3860–3862 (2000)
12. Yasumura, K.Y., Stowe, T.D., Chow, E.M., Pfafman, T., Kenny, T.W., Stipe, B.C., Rugar, D.: Quality factors in micro- and submicron-thick cantilevers. *J. Microelectromech. Syst.* **9**, 117–125 (2000)
13. Ibach, H.: Adsorbate-induced surface stress. *J. Vac. Sci. Technol. A* **12**, 2240–2245 (1994)
14. Ibach, H.: The role of surface stress in reconstruction, epitaxial growth and stabilization of mesoscopic structures. *Surf. Sci. Rep.* **29**, 193–263 (1997)
15. Grossmann, A., Erley, W., Hannon, J.B., Ibach, H.: Giant surface stress in heteroepitaxial films: invalidation of a classical rule in epitaxy. *Phys. Rev. Lett.* **77**, 127–130 (1996)
16. Nowick, A.S., Berry, B.S.: *Anelastic Relaxation in Crystalline Materials*. Academic, New York (1972)
17. Boland, J.J.: Structure of H-saturated Si(100) surface. *Phys. Rev. Lett.* **65**, 3325–3328 (1990)
18. Boland, J.J.: Role of bond-strain in the chemistry of hydrogen on the Si(100) surface. *Surf. Sci.* **261**, 17–28 (1992)
19. Wang, Y., Henry, J., Sengupta, D., Hines, M.: Methyl monolayers suppress mechanical energy dissipation in micromechanical silicon resonators. *Appl. Phys. Lett.* **85**, 5736–5738 (2004)
20. Henry, J., Wang, Y., Sengupta, D., Hines, M.: Understanding the effects of surface chemistry on q: mechanical energy dissipation in alkyl-terminated (c1–c18) micromechanical silicon resonators. *J. Phys. Chem. B* **111**, 88–94 (2007)

Surface Electronic Structure

Regina Ragan
Chemical Engineering and Materials, The Henry Samueli School of Engineering, Science
University of California, Irvine, CA, USA

Synonyms

[Local density of states](#)

Definition

Surface electronic structure is defined by filled and empty electronic states of the system near the surface of a solid material. In both bulk systems and nanosystems, the type of atoms in the system, the atomic arrangement, and atomic order versus disorder influence electronic structure.

Tight Binding, an early method for calculating electronic structure using a single particle Hamiltonian, approximates the electronic wave functions in a crystal as a linear combination of the atomic wave functions of constituent atoms. Tight Binding analysis shows that electrons in solid materials exhibit collective behavior that deviates from discrete electronic states in isolated atoms. Coupling of atomic wave functions in crystals leads to a continuum of energy states, called energy bands. The electronic structure at the surface often differs from that in the bulk due to broken bonds and atomic surface reconstructions. If the surface to volume ratio is high, as in nanosystems, then surface effects can dominate observed electronic properties.

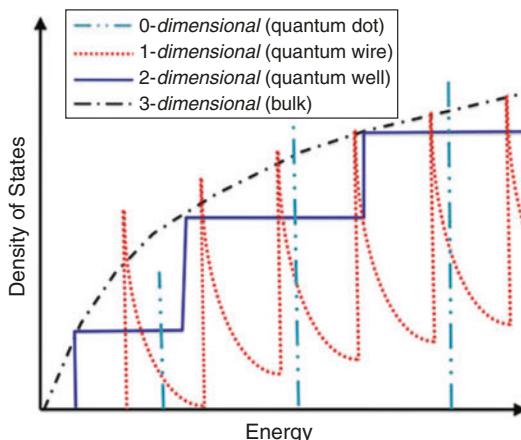
An important physical parameter to characterize electronic structure is the Fermi energy. The Fermi energy is often referred to as the highest occupied energy state since the probability of electron occupation of states having higher energy than the Fermi energy decays exponentially at equilibrium as determined from the Fermi-Dirac distribution function. The Fermi energy level is typically measured indirectly by the work function of the material. The work function is the energy needed to remove an electron from a material and is equal to the difference between the Fermi energy and the vacuum energy level. The position of the Fermi energy with respect to the energy bands determines whether the material is a metal, semiconductor, or insulator. Consider a simplified intuitive view. If atoms in the crystal have an odd number of valence electrons, then the energy bands that arise due to splitting of atomic energy levels are half filled. In this case, as found in metals, there is a continuum of available electronic states near the Fermi energy. A low resistance to electron motion is exhibited and thus metals are good conductors. In comparison, from this simple view, if there is an even number of valence electrons in constituent atoms of the crystal, then the (valence) energy bands are completely filled at zero Kelvin and the Fermi energy is in the energy bandgap. Near zero Kelvin, in both semiconductors and insulators, electrons do not occupy states in the conduction band and thus will exhibit negligible conductivity. Near

room temperature, semiconductors have few electrons at energy levels available for conduction.

In the case of nanosystems, the number of atoms in the system also affects electronic structure due to quantum confinement of electrons. In metals the “space” an electron occupies in the crystal is defined by the Fermi wavelength that has an inverse relationship with the electron density. In semiconductors, where charge carriers are electron and holes, the Bohr exciton radius defines the “space” for an electron hole pair. The Bohr exciton radius is a function of the dielectric constant of the material and the effective mass of the charge carriers. It is typically much larger than the Fermi wavelength in magnitude. When dimensions of nanoscale systems decrease below the Fermi wavelength (metals) or Bohr exciton radius (semiconductors) unique physical properties not observed in bulk materials arise due to quantum-size effects.

Overview

Drude theory provides a simple, intuitive analysis of the collective behavior of electrons in metallic systems. After J.J. Thompson discovered the electron in 1897, Drude treated valence electrons in metals as an electron gas and applied the kinetic theory of gases to describe electrical and thermal conductivity in metals. In this model, valence electrons are assumed to move freely in the solid and the potential exerted on electrons in the crystal lattice is assumed to be uniform due to nondirectional bonding of atoms in the crystal (metals have high coordination numbers). Electrons are only assumed to interact with ion cores during finite scattering events. Thus, the model is called the free electron model and the approximation for finite scattering events is called the relaxation time approximation. Although Drude theory only works well for alkali metals, it provides a qualitative understanding of thermal and electrical conductivity in metals. Early on Sommerfeld addressed some of the shortcomings of Drude theory to evaluate thermal conductivity by using Fermi-Dirac statistics to model the velocity distributions of electrons in metals in the context of



Surface Electronic Structure, Fig. 1 Schematic of density of states for three-dimensional, bulk system (black dot-dashed curve), two-dimensional, quantum well (blue solid curve), one-dimensional, quantum wire (red dashed curve), zero-dimensional, quantum dot (cyan double dot-dashed curve)

Drude theory. Quantum mechanical corrections are needed to accurately estimate the number of electrons that contribute to conductivity in most metals. Despite the need for these corrections, the approach of the electron gas to model observed electrical, thermal, and optical properties in materials was pivotal in our understanding of experimental observations in metals and semiconductors. The density of states, that is the number of available states per unit energy at a particular energy level, is derived from the electron gas approach. A schematic of how the density of states varies between bulk systems and quantum-confined systems is shown in Fig. 1. Van Hove singularities (discontinuity in the density of states) can be seen in quantum-confined systems. The concept of an electron gas is still used and describes quantum-confined systems such as metallic, single-walled carbon nanotubes (*one-dimensional* electron gas) that exhibit Van Hove singularities [1] and graphene (*two-dimensional* electron gas) that exhibits the quantum hall effect where the Hall conductivity exhibits quantized values [2].

When analyzing systems of atoms that do not form metallic bonds, ionic, and covalent systems, the approximations that bonding is nondirectional and the potential is uniform is no longer valid.

Kronig and Penny provided early intuition of energy levels and bands in crystals using both Bloch's theorem and a simple one-dimensional, periodic square wave potential that roughly approximates the potential of atoms in a periodic crystalline system. Via this simple analysis it is found that there are energy levels that yield nonphysical solutions for the electron wave functions and thus are not allowed for electrons in the system. These forbidden energy levels are defined as an energy bandgap. If the Fermi energy is in the bandgap, there are few electrons occupying energy states in the conduction band (above the Fermi energy) and thus conductivity is lower than in metals, hence the name semiconductor. As mentioned prior, the main difference between metals and semiconductors/insulators is that the Fermi energy sits within a band in a metal and in the energy bandgap in a semiconductor/insulator. The main difference between semiconductors and insulators is the magnitude of the bandgap. If greater than approximately 4 eV, then the probability of electrons occupying states in the conduction band is negligible at room temperature.

Basic Methodology

Computational Approaches

Significant advances in computational abilities coupled with important fundamental physical simplifications have allowed for first principles, ab initio, calculations of electronic structures of many body systems more closely modeling solid macroscopic and nanosystems and thus increasing accuracy. Since the Schrodinger equation cannot be solved analytically for these many electron systems, it was pivotal that Hohenberg and Kohn proved that the ground state energy of a many body system is a unique function of the charge density distribution [3]. Charge density distributions have a relationship with the atomic structure and periodicity in the crystal and thus can be evaluated by the types of atoms in the crystal and the crystal structure (atomic arrangement). Later, using an exchange correlation potential acquired from the theory of a homogeneous electron gas, Kohn and Sham determined how a many

body system could be reduced to a single particle equation for input into the Schrodinger equation. This is referred to as the local density approximation [3].

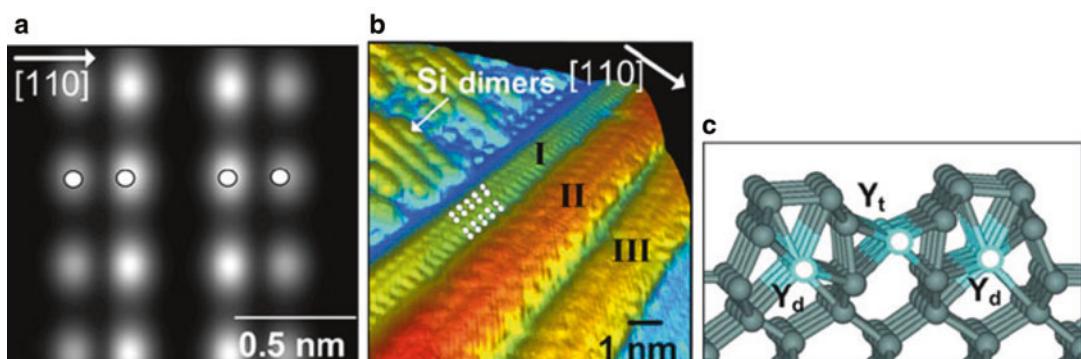
Due to these critical advances in understanding of many body systems, density functional theory (DFT) is widely used to calculate electronic structure that can be directly compared to experimental systems for understanding of experimental observations or used to predict physical properties to guide experiments. For example, DFT provided fundamental understanding of low-dimensional magnetism such as giant magnetic moments found in two-dimensional systems, such as monolayers of Mn and Cr. Giant magnetic moments arise due to an increase in the density of electronic states near the Fermi energy [3]. The phenomenon of Giant Magnetoresistance allowed for an increase in storage density in magnetic hard disk drives. Furthermore, low-dimensional metallic nanowires and nanoparticles can exhibit ballistic transport or unique chemical activity, respectively, and typically require feature sizes smaller than achievable with lithography. Thus self-organization is needed for fabrication. Atomic arrangements and driving forces for self-organization of low-dimensional metallic systems can be determined using atomic scale imaging in conjunction with DFT [4, 5]. Figure 2 shows how the correlation between DFT simulations and scanning tunneling microscopy (STM)

measurements allowed for an understanding of the atomic structure in disilicide nanowires that exhibit *one-dimensional* electron transport [4, 5]. The correct atomic arrangement is also critical for understanding the charge density distribution and in turn the surface electronic structure that is relevant, for example, to the development of nanocatalysts [6]. Improving efficiency and selectivity of catalysts will have significant economic benefit for the chemical industry. DFT has also been used to design electrode materials for lithium ion batteries. For example, Meng et al. have correlated structure with performance in electrodes using ab initio methods [7]. Materials design for batteries is crucial for development of plug-in electric vehicles that meet consumer performance requirements. Overall DFT allows for fundamental understanding of the relationship between atomic structure and electronic structure that is critical for understanding and utilizing physical properties of nanoscale systems.

Measurement Techniques

Photoelectron Spectroscopy

Photoelectron spectroscopy (PES) is a traditional surface analysis technique that uses a photon beam to provide energy for electrons to escape the potential of atoms or molecules near the surface. Kinetic energy distributions of emitted



Surface Electronic Structure, Fig. 2 (a) DFT simulation of the calculated electronic structure on the surface of an yttrium disilicide nanowire having a width of 1.1 nm. (b) STM image of dysprosium disilicide nanowires on Si (001) substrate. Nanowire labeled I has a width of 1.1 nm,

the calculated surface electronic structure matches the STM image as indicated by the white round circles. (c) Cross-sectional view of the relaxed atomic structure in the disilicide nanowires determined from DFT calculations and STM images (Printed with permission from Ref. [4])

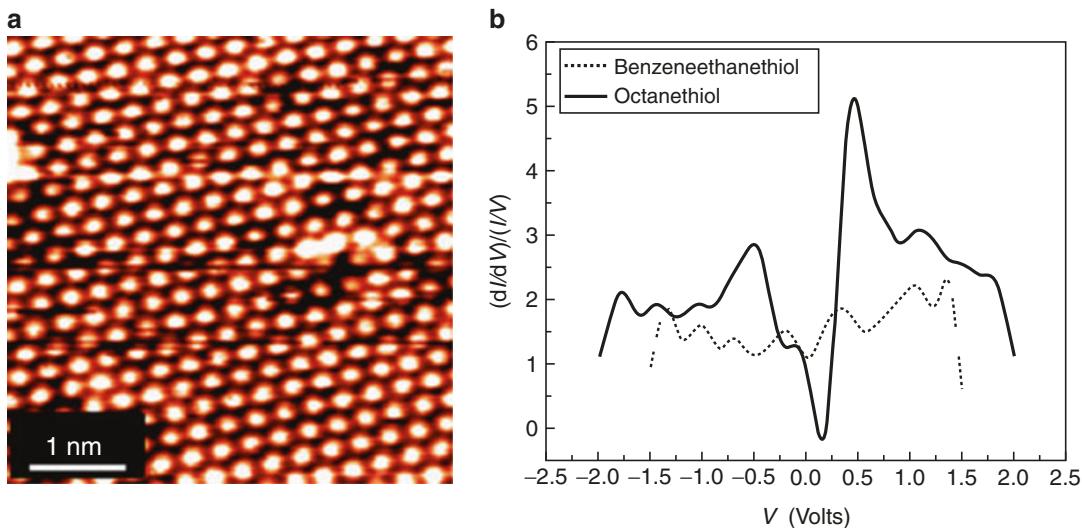
photoelectrons provide information regarding ionization energies or work function of the surface that reflects the composition and electronic states on the surface. Typical methods are x-ray photoelectron spectroscopy (XPS) that uses soft X-rays with energy on the order of 200–2,000 eV and ultraviolet photoelectron spectroscopy (UPS) that uses ultraviolet light with energy in the range of 10–45 eV. The energy of the photon beam affects the type of electrons that are emitted from the sample. During XPS core electrons are emitted from the atoms and during UPS the energy is sufficient to emit only valence electrons. Angle resolved photoemission spectroscopy (ARPES) provides additional information about the momentum of the emitted electrons. Since momentum is conserved, one can obtain the energy-momentum relationship (called the dispersion relationship) of the electrons in the crystal that reflects the energy band structure as a function of crystallographic direction. ARPES has been an important tool to measure the electronic band structure in bulk materials and in quantum-confined systems. For example, Yeom et al. measured a charge density wave in linear chains of indium atoms on silicon (001) surfaces [8]. Charge density waves are a coupling between electrons and lattice vibrations that exhibit a periodic modulation of charge that can be observed on a *one-dimensional* metallic surface. In the same study, a measured temperature dependent metal to semiconductor transition for in atomic chains on silicon (001) was attributed to a Peierls transition that is another signature of *one-dimensional* quantum confinement.

Scanning Probe Microscopy

Scanning probe microscopy (SPM) includes techniques such as scanning tunneling microscopy/spectroscopy (STM/STS), atomic force microscopy (AFM), electrostatic force microscopy (EFM), and Kelvin probe force microscopy (KPFM) that can be used to measure atomic structure and/or surface electronic structure. STM, invented in 1981 by Binnig and Rohrer, probes both atomic and electronic structure on surfaces by measuring tunneling current from the probe tip to sample surface across a narrow vacuum

(dielectric) gap. STM measurements typically achieve atomic scale spatial resolution. The AFM was invented shortly after the STM, 1986, to measure topography on nonconducting surfaces and is also capable of atomic and molecular resolution on surfaces. AFM measures van der Waals and electrostatic forces between cantilever tip and surface and thus does not require a conductive sample surface. KPFM is a variant of AFM in which conducting tips are used; KPFM allows for determination of the local surface potential with nanometer spatial resolution.

Tunneling current, as measured in STM, is extremely sensitive to both electron density and surface topography and convolutes the two properties. The tip-sample polarity affects whether electrons tunnel from sample to tip or from tip to sample and thus determines if the measurement probes filled or empty surface electronic states, respectively. A combination or empty and filled states imaging is often used in order to deconvolute the surface electronic structure from the atomic structure. STM can resolve corrugations heights on the sub-angstrom level, i.e., 0.2 Å corrugations between atoms on a clean platinum surface have been measured. The atomic arrangement on clean Pt(111) and how the surface structure evolves after depositing self-assembled monolayers on the surface have been measured using STM [9]. Figure 3a shows an STM image of Pt(111) where the hexagonal close packing of the atoms on the surface is easily observed. Scanning tunneling spectroscopy (STS) is a derivative of STM that measures the density of electronic states in the vicinity of the atomic scale STM tip, again yielding high spatial resolution. During STS, the current–voltage spectrum is measured between tip and surface while the voltage is swept across a specified range usually at constant tip-to-sample distance. Normalization of the differential conductance (dI/dV) with the conductance (I/V) can reflect the local density of states near the Fermi energy. Figure 3b shows normalized differential conductance data for benzene ethanethiol and octanethiol self-assembled monolayers on Pt (111) determined from STS data. Note that the Fermi energy is set at zero volts on the x -axis. The octanethiol/Pt(111) junction has zero



Surface Electronic Structure, Fig. 3 (a) STM image of Pt(111) surface. *R. Ragan unpublished results* (b) Normalized differential conductance of self-assembled

monolayers of benzenethiol (dashed curve) and octanethiol (solid curve) on Pt(111) (Printed with permission from Ref. [10])

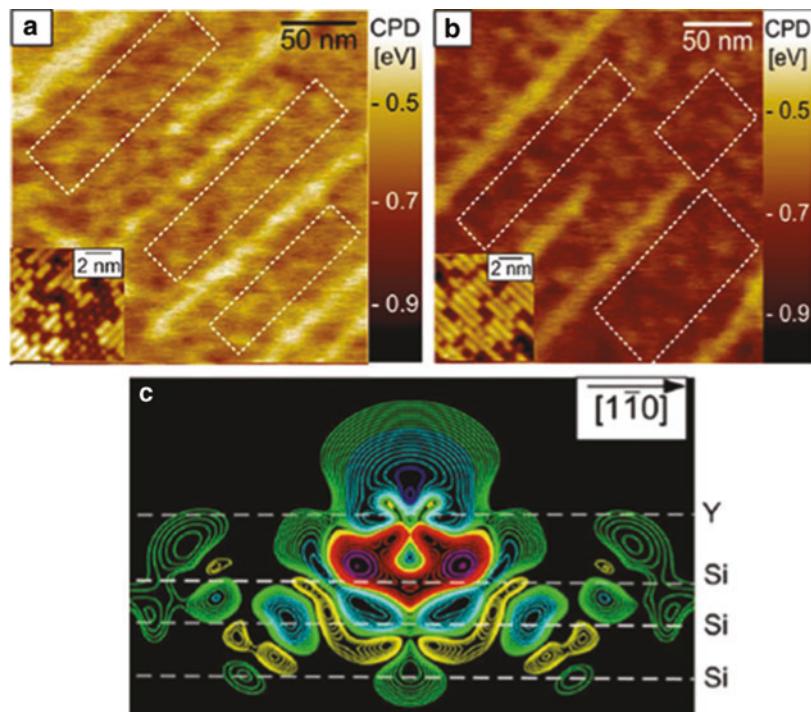
conductance at the Fermi energy and this represents an energy bandgap, i.e., no available states near the Fermi energy. For the benzene ethanethiol/Pt(111) junction, there is finite conductance at the Fermi energy and this is an indication of metallic behavior. A higher conductivity near the Fermi energy across benzene ethanethiol/Pt(111) junctions in comparison to octanethiol/Pt(111) junctions can be attributed to the fact that the benzene ethanethiol molecule has conjugated bonds in the molecule [9].

AFM measures the forces between AFM tip and surface by optically measuring the deflection of an AFM cantilever. With this basic mechanism, the type of signal feedback provides a wealth of information about the sample surface. Amplitude and frequency modulated AFM monitors the variation of the amplitude or frequency, respectively, of the AFM cantilever in response to forces between the tip and the surface and measures topography of the surface. In intermittent contact mode, the phase shift of the free cantilever resonance frequency provides nanometer scale information of the viscoelastic properties and adhesion force of the surface since it represents energy dissipation between tip and surface. For example, when imaging under a repulsive tip-sample

condition, regions of the surface with the higher elastic modulus appears darker in a phase contrast AFM image. One can observe a contrast reversal when the tip changes from repulsive to attractive mode. Variations in local topography also induce a phase shift in the cantilever frequency and thus topography and phase images need to be analyzed in conjunction to understand the physical properties of the surface.

Derivatives of AFM can be used to measure electrical properties when using a conducting cantilever. EFM measures electrostatic forces between surface and cantilever and is modeled by treating the vacuum, air or any dielectric gap between tip and surface as a capacitor. The force is dependent on the tip-surface distance and the potential difference between tip and surface. In particular, KPFM directly measures the contact potential difference (difference between sample surface work function and tip work function) using a lock-in technique to null the electrostatic forces between tip and sample surface. Recently, Ragan and Wu et al. have demonstrated that measured values of work function obtained from KPFM compare quantitatively with DFT calculations and together provide information on the atomic arrangement and termination of atoms on

Surface Electronic Structure, Fig. 4 KPFM images of dysprosium nanowires on Si(001) that have been annealed post-growth at (a) 600 °C and (b) 680 °C. (c) Simulated charge density difference image for a single metal adatom on Si(001) with the cross section perpendicular to the surface. Charge accumulation increases from yellow to pink contour lines, whereas charge depletion increases from green to blue contour lines. The greatest charge depletion is seen at the metal adatom location and the greatest charge accumulation is seen in the region between the adatom and the subsurface Si atoms (Printed with permission from Ref. [11])



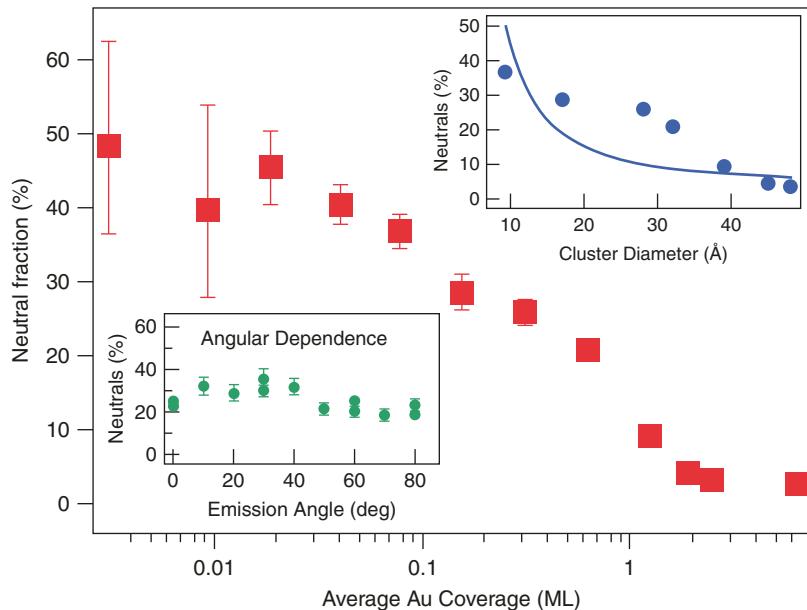
alloy surfaces [6]. Work function variations on surfaces reflect charge transfer [11, 12], quantum-size effects [13], and localized surface charge [14]. Figure 4 demonstrates how the work function on silicon changes when metal adatoms are on the semiconductor surface due to charge transfer between metal adatoms and substrate atoms. Figure 4a is a KPFM image after deposition of dysprosium on Si(001) and annealing the sample at 600 °C. Metallic disilicide nanowires form on the surface and appear as bright lines in the KPFM image since the disilicide nanowires have a lower work function than the Si(001) substrate. The Si(001) substrate in between nanowires is highlighted with white dashed boxes. A comparison of the substrate regions in Fig. 4a with those in Fig. 4b shows that the work function is lower on the substrate regions of the former. The STM images shown as insets in the lower left corner show that the Si(001) surface in Fig. 4a has more metal adatoms (dark regions in STM image) on the surface than the Si(001) surface of Fig. 4b due to the lower annealing temperature. The DFT simulation of Fig. 4c demonstrates that metal adatoms transfer charge to the Si atoms

on the surface. This creates a dipole on the surface that lowers the work function [11]. Overall, KPFM measurements provide information about material behavior in devices; KPFM across heterojunctions provides information regarding band offsets, device performance of diodes and chemical sensitive field effect transistors and trapped charge at interfaces in high electron mobility transistors.

Ion Scattering

Low energy ion scattering (LEIS) is a method complimentary to scanning probe techniques that also provides information about surface composition, electronic structure, and atomic structure. LEIS has less stringent requirements on surface conditions than SPM as data can be acquired using LEIS from rough and contaminated materials. The energy of scattered ions depends on the ratio of the projectile and target masses, providing a measure of the atomic mass distribution on the surface. The degree of charge exchange that occurs during scattering is dependent on the surface electronic properties when using projectiles with low ionization energies, such as alkalis.

Surface Electronic Structure, Fig. 5 Neutral fractions (NF) of singly scattered 2.0 keV Na ions shown as a function of the average Au coverage. The right side inset shows NF versus cluster diameter, with the symbols indicating experimental data and the solid line a theoretical fit. The left side inset shows NF for Na⁺ scattered from a 0.15 ML Au coverage as a function of the emission angle with respect to the surface normal (Printed with permission from Ref. [16])



When an alkali-metal atomic particle is in the vicinity of a surface, its ionization level shifts up due to the image charge interaction, while it broadens due to overlap of the ion and surface wave functions. The measured neutral fraction depends on the ionization potential, the degree that the level shifts near the surface, and the work function at a point just above the scattering site. The charge exchange process is well described by a non-adiabatic resonant charge transfer model [15]. Since the interaction is local to the site where projectile atoms exit the surface, the neutralization of low energy alkali ions provides a unique method for measuring the local work function and quantum-size behavior in nanomaterials.

Yarmoff et al. has used LEIS to measure quantum-size effects in the electronic structure of gold nanoclusters on TiO₂ [16] to provide insight regarding how catalytic activity and surface electronic structure are correlated. Au or other heavy metal nanocrystals are ideal for ion scattering experiments, as the large mass of the cluster atoms enables a complete separation of the ions that impact the nanoclusters from those that impact the substrate. The integrated single scattering peaks in “Neutrals” and “Total Yield” spectra are divided to obtain the neutral fraction for scattering from the

clusters. The neutral fraction for Na⁺ ions scattered from Au nanocrystals as a function of Au coverage is directly correlated with the size of the Au nanocrystals. The neutral fraction goes from about 50 % for the smallest clusters down to about 3 % for the film [16]. The enhanced neutralization from small clusters is due to participation of quantum-confined states in the non-adiabatic resonant charge transfer process. Bulk Au has a relatively high work function (5.1 eV), so that the Fermi level is degenerated with the Na ionization level (also \sim 5.1 eV) and most of the scattered Na remains ionic. The neutral fraction for small Au clusters is considerably larger than for bulk Au because filled states associated with the clusters provide electrons that can tunnel to the outgoing projectile. The existence of such filled states is consistent with reports that small Au clusters are negatively charged. The additional filled states above the Fermi level depend on the size of the clusters, and thus provide a measure of the quantum-size behavior (Fig. 5).

Examples of Application

Nanowire Sensors

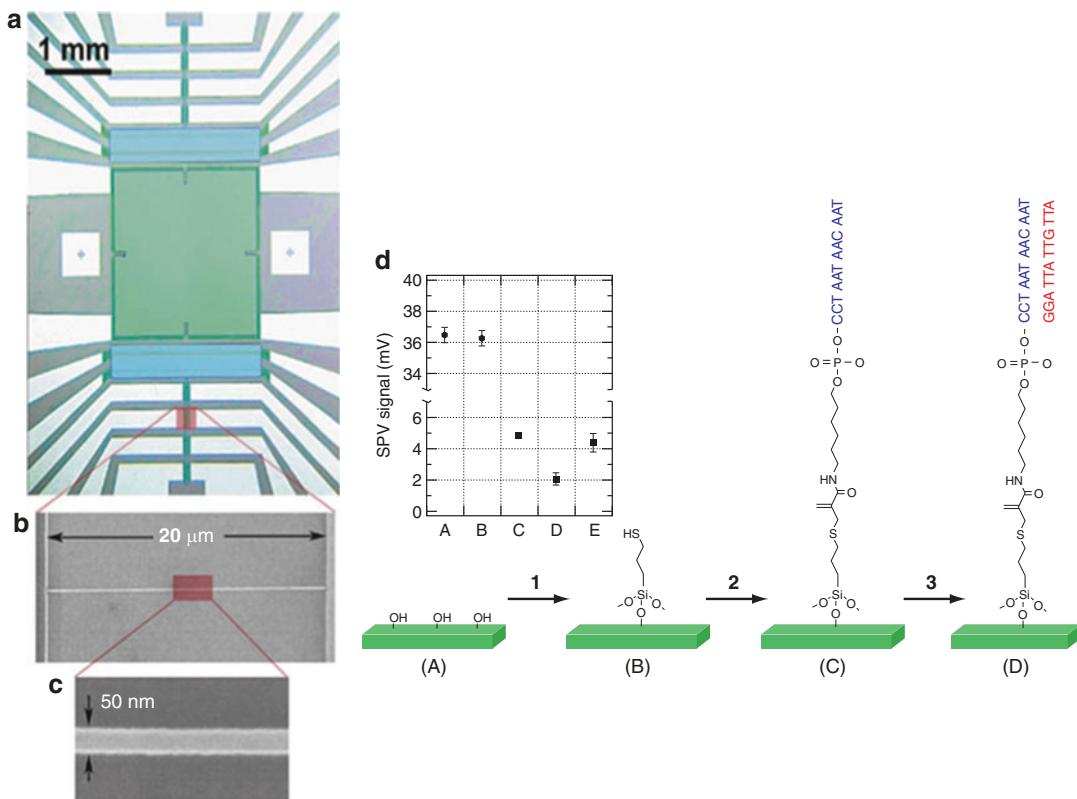
Chemical and biological nanowire sensors that use electronic transduction are an example of a

class of devices that uses changes in surface electronic structure to measure local molecular binding events or adsorption of molecules on surfaces. A chemical or biological sensor can be reduced to some basic components: a receptor (biological recognition element if selectivity is required), a transducer, and a means for processing whether or not target molecules of interest have bound to the surface and relaying this information to the user. Transducers can sense molecular interactions electronically (e.g., changes in localized charge) or optically (e.g., changes in dielectric constant). Changes in surface electronic structure can affect the conductance and this is easily measured in an electronic platform. A high surface to volume ratio in nanowires leads to high sensitivity for measuring changes in conductance due to surface binding events. If the surface of nanowires is functionalized with receptors that selectively bind to the target molecule of interest then selectivity can be engineered into the sensor platform. Since an early demonstration of semiconducting carbon nanotubes as chemical sensors, many different materials that form nanowires have exhibited the capability to sense molecules. In_2O_3 nanowires, SnO_2 nanoribbons, Si and ZnO nanowires have all been used as chemical and/or biological sensors. Si nanowires have been used for the detection of DNA hybridization, protein–protein interactions, cancer markers, and viruses. Some recent review articles that discuss nanowires-based sensor devices are referenced here [17, 18].

Nanowire sensors based on electronic transduction are commonly fabricated in a field effect transistor (FET) architecture with a back gate configuration. A single or multiple nanowires are connected to a source and drain and nanowires serve as the conducting channel. The conductance in the channel will change as molecules bind to the surface. Molecular binding events on semiconducting nanowire surfaces induce accumulation (depletion) of charge carriers that can be measured as an increase (decrease) lateral conductivity along nanowires in the FET architecture. The onset of accumulation or depletion of carriers depends on the surface charge induced and the carrier-type in the nanowire, *n-type* signifies that

the majority carriers are negatively charged electrons and *p-type* signifies that the majority carriers are positively charged holes. Thus in a FET architecture, molecular binding events on surfaces can be monitored in an electronic circuit. For example, in a study by Cui et al. Si nanowires were functionalized with biotin receptors. When streptavidin, target molecule, was introduced into the system at a concentration of 250 nM an increase in conductance was observed. Streptavidin is well known to have a high binding affinity with biotin. The increase in conductance due to the introduction of streptavidin could be measured for concentrations as low as 25 pM [17, 18]. This early study demonstrated both selectivity and high sensitivity of Si nanowire sensors.

Si has some advantages as a sensing platform since microelectronic circuits are typically made from Si. This approach provides a strategy for fabricating high-density, high-quality nanoscale sensors that can be integrated with Si-based circuits. Si nanowire sensors have been fabricated by either “top-down” or “bottom-up” fabrication methods where the former uses lithographic methods and the latter uses self-organization routes. Previously, Li et al. reported an approach to configure FET device architectures using Si nanowires for DNA sensors using a standard “top-down” semiconductor process [19]. Reactive ion etching is used to fabricate Si nanowires on silicon-on-insulator substrates that have been patterned using electron beam lithography. Figure 6a is an optical image of the entire sensor system. Figure 6b, c are high-resolution SEM images showing a Si nanowire in the system. The chemical functionalization process to bind a single strand DNA receptors on nanowire surfaces is illustrated in the schematic of Fig. 6d. The chemical modification process can be monitored after each step by measuring changes in surface potential using the surface photovoltage technique. When solutions containing complementary strands of DNA with concentrations of 25 pmol are introduced to *p-type* Si nanowire surfaces, there is an increase in conductance due to accumulation of carriers. In the case of *n-type* Si nanowires, the DNA binding event leads to a



Surface Electronic Structure, Fig. 6 (a) Optical image of lab on chip system using Si nanowires for signal transduction. (b) SEM image of Si nanowire spanning electrical leads. (c) High-resolution SEM image of Si nanowire. (d) Chemical functionalization of Si nanowire surfaces, steps

labeled (a–d), for sequence specific detection of DNA. The inset shows how the change in surface photovoltage (*SPV*) signal monitors each chemical functionalization step (Printed with permission from Ref. [19])

decrease in conductance due to depletion of carriers. Accumulation (*p*-type) or depletion (*n*-type) of carriers is associated with the negative charge on the backbone of DNA molecules. Quitoriano et al. introduced a different fabrication method for Si nanowire FET devices. This method is a combined “bottom-up” and “top-down” fabrication process utilizing the vapor–liquid–solid growth mechanism and optical lithography to fabricate FET devices [20, 21]. In this method, Au nanoparticles are deposited on the bottom surfaces of lithographically defined Si electrodes that overhang over a recessed trench of silicon dioxide. The growth of Si nanowires is guided from the bottom of the electrode at the site of the Au catalyst along the silicon dioxide layer to the opposite electrode. A benefit of combining

“bottom-up” methods with “top-down” methods is to integrate nanowires into microelectronic circuits using high-throughput methods. In both cases, Si nanowires sensors can be connected directly to the adjoining circuitry for signal amplification and automated data acquisition.

Nanoscale Catalysts

Metallic and bimetallic surfaces and nanostructures with tunable physical and chemical properties have attracted particular attention in recent years due to their potential for use in a broad range of applications. The discovery by Haruta, et al. that gold nanoparticles on oxide supports exhibit surprisingly high catalytic activity for reactions such as CO oxidation and propylene epoxidation has inspired an enormous wave

of research in the quest for innovative nanocatalysts. Chemically active Au nanoparticles have been prepared on reducible (TiO_2 , ZrO_2 , NiO , or Fe_2O_3) oxides, and irreducible (SiO_2 or Al_2O_3) oxides in order to gain insight on the role of quantum-size effects, facets and steps at the edges of nanoclusters, and charge transfer to the substrate in chemical activity. A review of mechanisms involved in enhanced activity of Au nanoclusters can be found here [22]. Bimetallic surfaces of Pt and Pd, and trimetallic nanoparticles of Au/Pt/Rh were also found to be highly effective in promoting a variety of reactions, typically higher than corresponding monometallic nanoparticles. Thus there appears to be a variety of parameters affecting chemical activity.

Surface electronic structure has been identified as playing an important role in enhanced catalytic activity in nanoscale metallic systems. Electronic structure can be modified in nanosystems in many ways such as by charge transfer between nanocatalyst and substrate support or due to alloy effects in bimetallic systems. Both experimental and theoretical results show the strong coupling between electronic structure and catalytic activity. The Hammer-Nørskov model predicted that chemisorption energy correlates with the d-band center in transition metals and this is experimentally observed for oxygen and sulfur chemisorption energies on different metal surfaces [23]. The effect of electronic structure and its relationship to catalytic activity is also observed in experiments. For example, Au clusters on TiO_2 were measured by STS to undergo a metal to insulator transition at nanometer length scales. The nanocluster size where the metal to insulator transition occurs exhibited the highest turnover frequency for CO oxidation, while larger clusters having no band gap had lower activity [24]. It has also been shown that catalytic reaction rates, in the context of electrochemical catalysis, exhibit an exponential dependence on the catalyst work function and catalytic rate enhancements of up to a factor of 60 having been reported. Changes in work function as small as 200 meV can lead to an increase in rate enhancement by a factor of 10 [25]. The work function of Au nanowires can

be varied by surface alloying [6] in order to optimize catalytic properties. Heterogeneous metal nanocatalysts with clusters that are a few nanometers in size hold great promise because of their large surface area to volume ratios, the availability of an enormous number of active sites and their enhanced resistance to poisoning from products of the reactions.

Cross-References

- [Ab Initio DFT Simulations of Nanostructures](#)
- [Atomic Force Microscopy](#)
- [Kelvin Probe Force Microscopy](#)
- [Nanomaterials for Electrical Energy Storage Devices](#)
- [Nanostructure Field Effect Transistor Biosensors](#)
- [Scanning Tunneling Microscopy](#)
- [Scanning Tunneling Spectroscopy](#)
- [Self-Assembly](#)

References

1. Wildoer, J.W.G., Venema, L.C., Rinzler, A.G., Smalley, R.E., Dekker, C.: Electronic structure of atomically resolved carbon nanotubes. *Nature* **391**, 59 (1998)
2. Berger, C., Song, Z.M., Li, T.B., Li, X.B., Ogbazghi, A.Y., Feng, R., Dai, Z.T., Marchenkov, A.N., Conrad, E.H., First, P.N., de Heer, W.A.: Ultrathin epitaxial graphite: 2D electron gas properties and a route toward graphene-based nanoelectronics. *J. Phys. Chem. B* **108**, 19912 (2004)
3. Freeman, A.J., Wu, R.Q.: Electronic-structure theory of surface, interface and thin-film magnetism. *J. Magn. Magn. Mater.* **100**, 497 (1991)
4. Shinde, A., Wu, R., Ragan, R.: Thermodynamic driving forces governing assembly of disilicide nanowires. *Surf. Sci.* **604**, 1481 (2010)
5. Zeng, C., Kent, P.R.C., Kim, T., Li, A., Weitering, H. H.: Charge-order fluctuations in one-dimensional silicides. *Nat. Mater.* **7**, 539 (2008)
6. Ouyang, W., Shinde, A., Zhang, Y., Cao, J., Ragan, R., Wu, R.: Structural and chemical properties of gold rare earth disilicide core – shell nanowires. *ACS Nano* **5**, 477 (2011)
7. Meng, Y.S., Arroyo-de Dompablo, M.E.: First principles computational materials design for energy storage materials in lithium ion batteries. *Energy Environ. Sci.* **2**, 589 (2009)

8. Yeom, H.W., Takeda, S., Rotenberg, E., Matsuda, I., Horikoshi, K., Schaefer, J., Lee, C.M., Kevan, S.D., Ohta, T., Nagao, T., Hasegawa, S.: Instability and charge density wave of metallic quantum chains on a silicon surface. *Phys. Rev. Lett.* **82**, 4898 (1999)
9. Ragan, R., Ohlberg, D., Blackstock, J.J., Kim, S., Williams, R.S.: Atomic surface structure of UHV-prepared template-stripped platinum and single-crystal platinum(111). *J. Phys. Chem. B* **108**, 20187 (2004)
10. Lee, S., Park, J., Ragan, R., Kim, S., Lee, Z., Lim, D. K., Ohlberg, D.A.A., Williams, R.S.: Self-assembled monolayers on Pt(111): molecular packing structure and strain effects observed by scanning tunneling microscopy. *J. Am. Chem. Soc.* **128**, 5745 (2006)
11. Shinde, A., Cao, J.X., Lee, S.Y., Wu, R.Q., Ragan, R.: An atomistic view of structural and electronic properties of rare earth ensembles on Si(001) substrates. *Chem. Phys. Lett.* **466**, 159 (2008)
12. He, T., Ding, H.J., Peor, N., Lu, M., Corley, D.A., Chen, B., Ofir, Y., Gao, Y.L., Yitzchaik, S., Tour, J. M.: Silicon/molecule interfacial electronic modifications. *J. Am. Chem. Soc.* **130**, 1699 (2008)
13. Lee, S., Shinde, A., Ragan, R.: Morphological work function dependence of rare-earth disilicide metal nanostructures. *Nanotechnology* **20**, 6 (2009)
14. Rosenwaks, Y., Shikler, R., Glatzel, T., Sadewasser, S.: Kelvin probe force microscopy of semiconductor surface defects. *Phys. Rev. B* **70**, 085320 (2004)
15. Kimmel, G.A., Goodstein, D.M., Levine, Z.H., Cooper, B.H.: Local adsorbate-induced effects on dynamic charge-transfer in ion-surface interactions. *Phys. Rev. B* **43**, 9403 (1991)
16. Liu, G.F., Sroubek, Z., Yarmoff, J.A.: Detection of quantum confined states in Au nanoclusters by alkali ion scattering. *Phys. Rev. Lett.* **92**, 216801 (2004)
17. Patolsky, F., Lieber, C.M.: Nanowire nanosensor. *Mater. Today* **8**, 20 (2005)
18. Kolmakov, A., Moskovits, M.: Chemical sensing and catalysis by one-dimensional metal-oxide nanostructures. *Annu. Rev. Mater. Res.* **34**, 152 (2005)
19. Li, Z., Chen, Y., Li, X., Kamins, T.I., Nauka, K., Williams, R.S.: Sequence-specific label-free DNA sensors based on silicon nanowires. *Nano Lett.* **4**, 245 (2004)
20. Quitoriano, N.J., Kamins, T.I.: Integratable nanowire transistors. *Nano Lett.* **8**, 4410 (2008)
21. Quitoriano, N.J., Wu, W., Kamins, T.I.: Guiding vapor-liquid-solid nanowire growth using SiO₂. *Nanotechnology* **20**, 145303 (2009)
22. Min, B.K., Friend, C.M.: Heterogeneous gold-based catalysis for green chemistry: low-temperature CO oxidation and propene oxidation. *Chem. Rev.* **107**, 2709 (2007)
23. Greeley, J., Norskov, J.K., Mavrikakis, M.: Electronic structure and catalysis on metal surfaces. *Annu. Rev. Phys. Chem.* **53**, 319 (2002)
24. Valden, M., Lai, X., Goodman, D.W.: Onset of catalytic activity of gold clusters on titania with the appearance of nonmetallic properties. *Science* **281**, 1647 (1998)
25. Vayenas, C.G., Bebelis, S., Ladas, S.: Dependence of catalytic rates on catalyst work function. *Nature* **343**, 625 (1990)

Surface Energy and Chemical Potential at Nanoscale

Vanni Lughì

DI3 – Department of Industrial Engineering and Information Technology, University of Trieste, Trieste, Italy

Synonyms

Interfacial energy and chemical potential at nanoscale; Nanothermodynamics; Surface energy density and chemical potential at nanoscale; Surface free energy and chemical potential at nanoscale; Surface tension and chemical potential at nanoscale

Definition

Chemical potential is the energy increment of a system, associated to the addition of one single element of the substance – e.g., one atom or molecule. *Surface energy* is commonly defined as the energy increment associated to the formation of a unit area of a new surface. Although the latter is by far the most common and practical definition of surface energy and serves its purpose in most physical situations, it is subject to a number of ambiguities, which will be sorted out in the following section.

In nanoscale systems, a large portion of the atoms or molecules constituting the system is at or near surfaces. Knowledge of the surface properties, and in particular of surface energy, is therefore a key step in understanding and controlling nanostructures and nanostructured materials. On the other hand, surface geometry affects the chemical potential of the system, as shall be seen below,

and this effect becomes most important at the nanoscale. Chemical potential and surface energy are therefore not only key elements when it comes to describing systems at the nanoscale, but they are also somewhat interrelated and are therefore presented here together.

In most of current studies and applications, the subjects of surface energy and chemical potential can be properly discussed within a classical approach – and this is done in most of this entry. The more consistent and modern, but less agile, approach of nanothermodynamics is briefly introduced in the last paragraph.

Fundamental Considerations on Surface Energy

Before turning to the specifics of the subject, it is important to sort out some ambiguities that normally arise when discussing surface energy. In Gibbs' approach, the surface of a solid or liquid is considered as a transition volume between the bulk and its vapor, as depicted in Fig. 1. This

model can in principle be generalized to multicomponent systems. All extensive thermodynamic quantities that characterize the system – such as volume, V ; number of atomic or molecular constituents, N ; internal energy, U ; and entropy, S – can therefore be partitioned and a specific portion V_S , N_S , U_S , and S_S is assigned to the “surface volume” (Fig. 1). As these quantities represent extra amounts with respect to the bulk values, they are known as *excess* quantities. By combining these surface excesses, the Gibbs free energy $G_s = U_S - TS_S + PV_S$ of the surface and the Helmholtz free energy $F_s = U_S - TS_S$ of the surface can be defined – in complete analogy with the bulk. G_s and F_s are also excess quantities. Here T and P are the temperature and the pressure, respectively. Without loss of generality [1, 2], it is possible to assign the excess quantities that originally pertain to the surface volume to a representative, two-dimensional interface of area A (dotted line in Fig. 1). In this case, the surface energies can finally be defined:

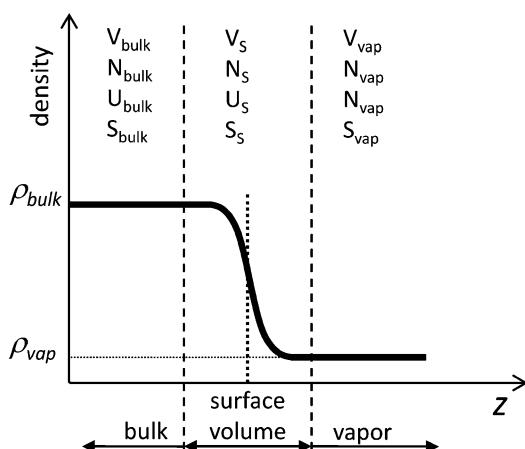
$$\text{Surface energy } u_s = \frac{U_s}{A} \quad (1)$$

$$\text{Gibbs surface free energy } g_s = \frac{G_s}{A} \quad (2)$$

$$\text{Helmholtz surface free energy } f_s = \frac{F_s}{A} \quad (3)$$

Hence, strictly speaking, surface energy is defined as the internal energy of the surface per unit area, and the surface free energy is defined as the free energy (Helmholtz or Gibbs) of the surface per unit area, where in this case the “surface” is defined as the transition volume between the semi-infinite, bulk phases of the system as illustrated in Fig. 1.

In a more common approach, one can consider the entire free energy, G , of the generic one-component solid system characterized by the presence of a surface [2, 3] (the equation can be generalized to multicomponent systems by adding a $\mu_i N_i$ term for each i th component):



Surface Energy and Chemical Potential at Nanoscale,
Fig. 1 Gibbs model of a solid in equilibrium with its vapor (one-component system). The density, ρ , of the component is plotted as a function of the distance normal to the surface, z . The volume partition and the associated extensive quantities are indicated by the dashed lines

$$G = U - TS + PV = \mu N + \gamma A \quad (4)$$

Here μ and γ are the intensive quantities *chemical potential* and *surface tension*, which are therefore thermodynamically defined as:

$$\gamma = \frac{\partial G}{\partial A} \Big|_{T, P, N} \quad (5)$$

$$\mu = -\frac{\partial G}{\partial A} \Big|_{T, P, N} \quad (6)$$

Adsorption and Relationship Between Surface Tension and Surface Energy

Unfortunately, in most materials science, nanoscience, and nanotechnology literature, the distinction between surface energy, surface free energy, and surface tension is not always maintained. Although strictly speaking γ as defined in Eq. 5 is the surface tension and can be used as such for solids and liquids, the terms “surface energy” and “surface free energy” are also commonly used to identify the same quantity. In general, however, surface tension (Eq. 5) and surface (free) energy (Eqs. 1, 2, and 3) are different, and the difference depends on the surface excess of species: This can be in general the absorption of foreign species, or for the case of multicomponent systems simply the surface excess per unit area of the i th component of the system $\Gamma_i = N_{S,i}/A$. As shown by Gibbs [1], one finds:

$$\gamma = f_s - \sum_i \Gamma_i \mu_{S,i} \quad (7)$$

Equation 7 can be used to relate the concentration of a solute at the surface with a change of the surface energy. In pure one-component systems the summation is nil and $\gamma = f_s$, but this is a rather ideal case. In common systems, it is virtually impossible to obtain perfectly clean surfaces. In nanosystems foreign species are often added on purpose – such as in the case of colloidal nanoparticles, where capping agents are added to avoid the formation of clusters.

In the following, the term “surface energy” will be used as a generic term indicating the relevant one among “surface tension” (Eq. 5) and the different surface free energies defined in Eqs. 1, 2, and 3. These are all intensive quantities. Throughout this entry, the term “energy of the surface” will be used to indicate the relevant one among the extensive quantities γA , $g_S A$, or $f_S A$, where A is the surface area of the system. Note that some authors prefer to use the term “surface (free) energy density” when referring to these intensive quantities, in order to avoid confusion with the extensive quantities.

Surface Stress

In the definition of surface energy given above, the implicit mechanism for the formation of a new surface is cleaving of a bulk material. However, the area of an existing surface can be increased by other mechanisms, i.e., by applying a deformation (stretching, distortion). In this case a term needs to be added to Eq. 4:

$$G = U - TS + PV = \mu N + \gamma A + \xi A \quad (8)$$

Here ξ is the surface deformation energy density. For an elastic solid it is $\frac{1}{2} \sum_{i,j} \sigma_{ij} \varepsilon_{ij}$ where σ_{ij} and ε_{ij} are the tensor components of the *surface stress* and *strain*, respectively, and are reciprocally correlated by the compliance properties of the material. (Note that in this case σ_{ij} and ε_{ij} are defined as surface tensors in complete analogy with the bulk, and have units of force per unit length.) Some mathematical manipulation (see for example [2]) leads to:

$$A dy + S_S dT + V_S dP + N_S d\mu + A \sum_{i,j} (\gamma \delta_{ij} - \sigma_{ij}) d\varepsilon_{ij} = 0 \quad (9)$$

where S_S , V_S , and N_S are the excess quantities as defined previously and δ_{ij} is the Dirac delta function. Equation 9 is one of the key results in surface thermodynamics: On one hand, it can be regarded as a general form of Eq. 7 (some literature reports it as the *Gibbs adsorption equation*); moreover, it shows the relationship between chemical potential, surface energy, and surface stress (note that,

thanks to the Gibbs-Duhem equation $SdT - VdP + Nd\mu = 0$, which defines the general relationship between the intensive thermodynamic parameters, of the five variables in Eq. 9 ($\gamma, \mu, \varepsilon, P, T$, only three are independent). Finally, from Eq. 9 it can be shown [2] that:

$$\sigma_{ij} = \gamma\delta_{ij} + \frac{\partial\gamma}{\partial\varepsilon_{ij}} \Big|_T \quad (10)$$

Equation 10 shows that surface stress is rigorously equal to surface tension only when there is no dependence of the latter on the surface deformation. This is normally only true in liquids, where the surface atoms can rapidly rearrange in response to a deformation, whereas in solids surface stresses have to be relieved by opportune mechanisms, such as dislocations or buckling of the surface. In solid nanosystems, due to the predominance of surfaces, these effects can have a rather large impact.

Typical Values

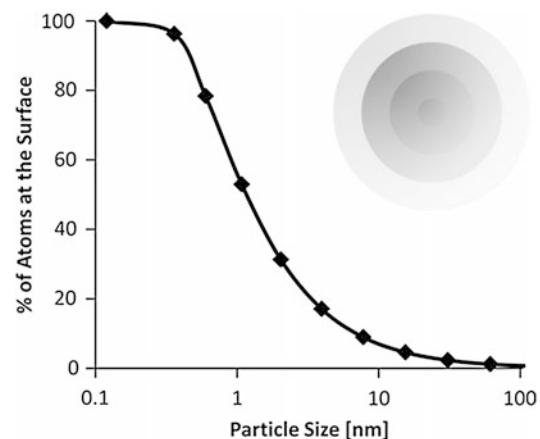
When a surface is created, a number of atomic bonds are broken. This consideration enables a practical estimation of the surface energy, since for a solid-vapor interface, the energy per unit area associated to the broken bonds is simply $\frac{1}{2}\varepsilon N_{bb}\rho_S$ where $\varepsilon/2$ is half of the bond energy, N_{bb} is the number of broken bonds per surface atom, and ρ_S is the planar atomic density at the surface (atoms per unit area). The bond energy can in turn be estimated – quite precisely in some instances such as the case of pure metals – from the latent heat of sublimation: $L_{sub} = \frac{1}{2}\varepsilon N_A N_{nn}$ where N_A is Avogadro constant and N_{nn} is the number of atom's nearest neighbors in the bulk [3]. L_{sub} is a known quantity that can be found in thermodynamic tables. The one described here, however, only provides a rough estimate. An entropic term should also be considered, and a number of structural rearrangements of the system, which shall be described in the following section, contribute to the actual, lower value of the surface energy.

Surface energy and surface tension are reported indistinctly with units of $J\ m^{-2}$ or $N\ m^{-1}$, although the latter is often preferred for

expressing surface tension in liquids. Common values range from a few tens (in liquids, in polymers, and in ceramics) to a few thousands (in metals) of $mJ\ m^{-2}$.

Surface Energy at the Nanoscale

Nanosystems have very large surface extensions with respect to bulk systems. Based on geometrical considerations, it is easy to show that once the total amount of material is fixed, the total surface area of an ensemble of objects is inversely proportional to the linear size of the object, and so will be the contribution of the energy of the surface to the total energy system. For instance, an ensemble of nanoparticles with 10 nm diameter has 10^6 times more surface area (and energy of the surface) than the same amount of material organized in a single 1-cm particle. From a different standpoint, one can observe that the fraction of atoms that lie at the surface increases at the nanoscale. Figure 2 in particular shows a very sharp increase when reducing the size below ~ 20 nm. Materials properties that are strongly dependent on the energy of the system are expected to undergo dramatic changes in a similar size range. A number of such effects are briefly



Surface Energy and Chemical Potential at Nanoscale, Fig. 2 Geometrical approximation of the fraction of atoms at the surface of a spherical particle as a function of the sphere diameter. The curve is constructed by approximating the number of surface atoms with the volume of concentric spherical shells (inset)

discussed in the next paragraph dedicated to the chemical potential.

Minimization Mechanisms for the Energy of the Surface: Role in Nanostructures

As part of the total energy minimization process, a system will in general undergo a number of rearrangements at all scales in order to reduce the energy of the surface [4]. In nanosystems, this is particularly important for two reasons: Firstly, because the energy of the surface constitutes a large portion of the total system's energy. Moreover, even small rearrangements that would hardly affect a macroscopic system's morphology and/or properties can strongly impact structures with nanometric size. There are three categories of mechanisms that can contribute to reducing the total energy of the surface in a system, involving:

- (a) local surface phenomena, which act on reducing the surface energy of the structures;
- (b) the individual structure;
- (c) the overall system.

These mechanisms have general validity, but will be discussed in the following from the standpoint of nanosystems, for which they are most relevant.

Local Surface Mechanisms. The first and unavoidable mechanism is *surface relaxation*: Atoms at the surface cannot maintain the position they would have in the bulk, as they would be subject to asymmetric forces, and are therefore bound to find a different equilibrium position. This normally results in an inward and/or lateral shifts of the near-surface atomic layers, depending on the crystal symmetry. Whereas in bulk materials the associated reduction of the average lattice constant is essentially negligible, it can be noticeable in some small nanoparticle systems. If there is more than one broken bond per atom, relaxation of the surface can occur by *reconstruction*, where new bonds form at the surface and the geometry of the surface changes. A classic case is the 7×7 reconstruction of clean (111) silicon surfaces. If foreign species are available, *physical or chemical adsorption* also reduces surface energy. In atmosphere, essentially all surfaces are covered by adsorbed species, most commonly hydrogen or hydroxyl groups. Finally, changes of the chemical composition at the surface, for instance by *segregation* of impurities, are another effective way to

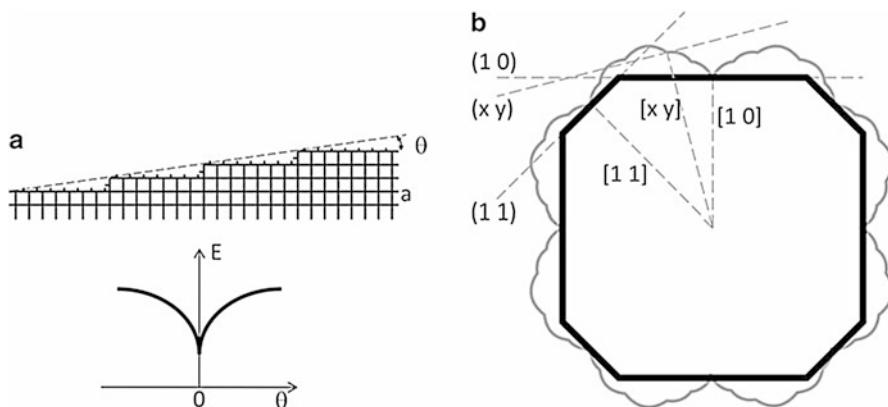
reduce surface energy. This mechanism can have a strong impact on the properties of nanoparticles and nanowires: In these cases segregation at the surface can fully deplete any impurity in the interior of the nanostructure. If on one hand this is one of the reasons for the high purity of nanocrystals, on the other it is a major, fundamental obstacle for achieving effective electronic doping in semiconducting nanostructures.

Individual Structure Mechanisms. An individual structure – e.g., nanoparticles, nanowires, etc. when considering the nanoscale – can reduce the energy of the surface by opportunely changing shape. In isotropic systems, such as liquids or amorphous solids, surface area, A and total energy of the surface, γA , are strictly proportional, therefore minimizing the former also minimizes the latter.

For anisotropic systems, such as crystals, the surface energy varies from one crystal plane to the other. The quantity that needs to be minimized is $\sum_i A_i \gamma_i$ (the index i refers to the different crystallographic planes in the crystal) so that minimization of the area is not in general the optimal solution. This problem is solved by using the *Wulff construction*, which predicts the equilibrium shape of crystals. As illustrated in Fig. 3a, a *vicinal surface* – i.e., a crystal plane forming a small angle θ with the close-packed plane – will have additional broken bonds with respect to a close-packed plane. The number of broken bonds on such a vicinal surface can be derived from simple geometrical considerations:

$$N_{bb} = \frac{\cos\theta + \sin|\theta|}{2a^2} \quad (11)$$

where a is the lattice parameter [2]. In proximity of a close-packed plane the surface energy will then have the shape shown in Fig. 3a, in accordance with Eq. 11: The close-packed plane, characterized in general by small Miller indexes, is associated to a local minimum. The surface energy for all crystal directions can be plotted in a polar diagram, as shown in Fig. 3b for a two-dimensional cross section. One can draw a radius vector for each crystal direction, and then,



Surface Energy and Chemical Potential at Nanoscale,
Fig. 3 (a) Schematic cross section of a vicinal surface forming an angle θ with a close-packed plane, and diagram of the surface energy as a function of the angle. (b) Example of a two-dimensional polar plot of the surface energy (gray solid line). The *dashed lines* indicate the Wulff

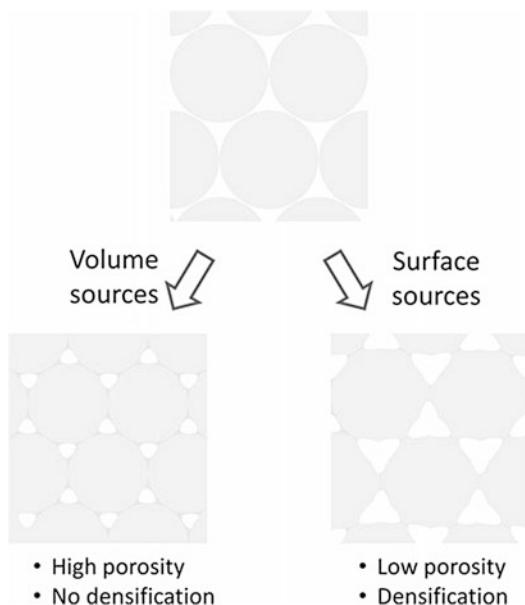
at the intersection point, the plane that is perpendicular to such vector. Wulff demonstrated that the internal envelope of all such planes defines the equilibrium shape of a crystal, as it minimizes the quantity $\sum_i A_i \gamma_i$. While in macroscopic crystals the shape predicted by the Wulff construction is rarely observed because the kinetics to equilibrium are rather slow at the temperatures of interest, nanosystems often do reach equilibrium due to the short spatial scales involved, and the approach described here can be a useful tool for predicting the shape of nanocrystals.

Mechanisms Involving the Overall System. A system consisting of an ensemble of structures can reduce the overall energy of the surface by simply *aggregating* such structures to one another, thus reducing the surface area. An important case is that of suspensions colloidal particles, where stabilization mechanisms (such as electrostatic charging, or steric stabilization by coating the particles with organic ligands) are needed especially for nanosized particles. Lack of proper stabilization leads to aggregation of the nanoparticles with the formation of large clusters, and the benefits of the colloidal suspension as well as the nanoscale properties of the particles are lost.

An aggregate, e.g., of particles, can further reduce the energy of the surface by *sintering*. During sintering, atoms move toward concave

construction of the planes perpendicular to selected crystallographic directions. The *black solid line* indicates the expected equilibrium shape of the crystal. Note that the generic (xy) does not contribute to the final equilibrium shape because of the high surface energy associated to it

surfaces driven by diffusion mechanisms, which in turn are controlled by the radius of curvature of the target surface: The lower the radius (negative values correspond to concave surfaces), the faster the diffusion toward that area. This results in the progressive merging of adjacent particles, as depicted in Fig. 4. The sources of atoms can be the interior volume of the particles, the surface of the particles, or the boundaries between particles (i.e., as sintering proceeds, the *grain boundaries*). When the source of atoms is the interior volume or a grain boundary, the particles get closer, and densification can occur. If properly conducted, sintering can lead to the formation of a fully dense solid from an aggregate of particles. When the source of atoms is the surface, then no densification can occur, and the final structure will have a high porosity – and therefore still a high surface area and energy of the surface. The dominant source of atoms strongly depends on the sintering conditions and can in principle be controlled, thus providing a tool for tailoring the material's morphology at the nanoscale and consequently engineering its properties: On one hand, obtaining a fully dense solid with grains of nanoscale size has generated quite a bit of interest because it leads to excellent mechanical properties and, in the case of materials subject to phase transformation, to enhanced phase stability (as illustrated in the



Surface Energy and Chemical Potential at Nanoscale,
Fig. 4 Schematic of a sintering process and of the resulting microstructures. (*Left*) If the source of the diffusing species is the volume of the particles or the grain boundaries, densification occurs: The distance between the particle centers decreases, and the porosity is reduced. (*Right*) If the source of the diffusing species is the particle surface, no densification can occur: The material is just redistributed along the surfaces, favoring the necks that are now forming between particles, and the particle centers do not change position

following paragraph). On the other hand, obtaining a highly porous material, where the pore size is at the nanoscale, is of interest for a number of situations including: high-performance gas sensors; separation of pollutants in environmental applications; catalysis and photocatalysis, including innovative routes for producing and storing hydrogen for energy applications; scaffolds in biological and biomedical applications; electrodes for fuel cells and batteries; high-performance thermal insulators. However, sintering in nanosystems can also be an unwanted effect. This is especially true when the nanosystem is subject to heat treatment, even at rather low temperatures: The reason is that the diffusion mechanisms that govern sintering become well active at temperatures close to 50–70 % of the melting temperature, which in

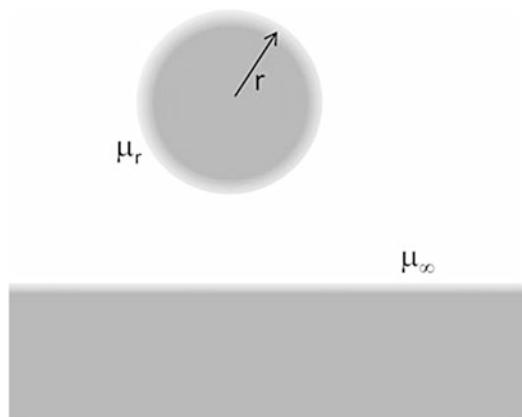
turn drops dramatically at the nanoscale as mentioned in the following paragraph.

Finally, *Ostwald ripening* is another important mechanism where interaction between the elements of the entire system leads to reduction of the overall surface area and of the energy associated to it. In this case, larger particles grow at the expenses of smaller particles. The reason is that atoms at the surface of a small particle are less stable than those at the surface of a larger particle. (More details on this mechanism will be given in the following section.) Ostwald ripening is of major importance when dealing with nanostructures, where small radii of curvature are common. Important examples include sintering processes of nanopowders – in particular densification processes, where the ripening should be avoided as it conflicts with the need of keeping grain size small and uniformly distributed. (See for example the phenomenon of “exaggerated grain growth” typical of ceramic materials processing.) Also, Ostwald ripening is often observed during the synthesis of colloidal nanoparticles – where it can either be deleterious or, if opportunely controlled, lead to a better size distribution.

Chemical Potential at the Nanoscale and Relationship with Surface Energy

The definition of chemical potential was given above as the energy increment of a system associated to the addition of one atom or molecule, and was expressed quantitatively in Eq. 6. One requirement that is explicit in this equation is that the area of the system must remain constant. In general, this is not true, as addition of even one atom or molecule to the material does change, in principle, the surface area. This effect is only important when the size of the system is small. It can be shown [3, 4] that the difference between the chemical potential, μ_r , of a spherical particle of radius r and the chemical potential, μ_∞ , of a semi-infinite bulk with a flat surface made of the same material (Fig. 5) is:

$$\mu_r - \mu_\infty = \frac{2\gamma\Omega}{r} \quad (12)$$



Surface Energy and Chemical Potential at Nanoscale,
Fig. 5 Model for calculating the chemical potential difference, $\mu_r - \mu_\infty$, between a particle of radius r and a semi-infinite bulk with a flat surface

where Ω is the atomic volume. This relationship, named after Gibbs and Thomson who first independently derived it, is obtained by realizing that the energy change due to the addition of a single atom to the particle, $\mu_r \partial N$, must be equal to the energy change associated to the change in surface area, $\gamma \partial A$, and mathematically manipulating the balance equation. The term $2\gamma/r$ is known as the Laplace pressure, and it can be shown that the pressure difference, Δp , between the material (solid or fluid) inside and outside of the curved surface is:

$$\Delta p = \frac{2\gamma}{r} \quad (13)$$

Known as the *Young-Laplace equation*, this is also often regarded as a mechanical definition of the surface tension, and is one of the key equations when studying micro- and nanofluidics. Note that, in the case of a solid, a consequence of this relationship is that in a small crystal, there should be a contraction of the lattice parameter, an effect that is actually observed but only for extremely small particles (below about 1 nm). Equations 12 and 13 can be extended to any curved surface, by substituting $2/r$ with the generalized local curvature of the surface, $(1/r_1 + 1/r_2)$, where r_1 and r_2 are the principal radii of curvature.

The Gibbs-Thomson relationship shows the coupling between surface energy and chemical potential, and provides a solid framework for deriving their combined effects on mechanisms and properties of systems, especially at the nanoscale where $\mu_r - \mu_\infty$ can be very large and the effects are most evident. Some important examples are briefly described in the following [4].

First, the Gibbs-Thomson can be extended to the *vapor pressure* of the system. Assuming an ideal gas behavior for the vapor, one can show the chemical potential of an atom in the vapor phase is $\mu_v = \mu_\infty - kT \ln P_\infty$ where P_∞ is the equilibrium vapor pressure of a flat surface, T the temperature, and k Boltzmann's constant; an analogous relationship holds when a curved surface is involved and P_r is the equilibrium vapor pressure of a curved surface. It is therefore possible to calculate $\mu_r - \mu_\infty$ and to use Eq. 12 to find the *Kelvin equation*:

$$\ln \frac{P_r}{P_\infty} = \frac{2}{r} \frac{\gamma \Omega}{kT} \quad (14)$$

The vapor pressure increases as the inverse of the particle (or droplet in the case of a liquid) radius. Once again, the effect becomes sensible at the nanoscale. An identical relationship can be derived for the *solubility* of solids, S , revealing that nanoparticles are much more soluble than the bulk counterpart:

$$\ln \frac{S_r}{S_\infty} = \frac{2}{r} \frac{\gamma \Omega}{kT} \quad (15)$$

The solubility difference between particles of different radius is also strictly correlated with *Ostwald ripening*, a phenomenon that is observed in colloidal suspensions of nanocrystals or during sintering processes of ceramic nanopowders. The driving force for this mechanism is the difference in chemical potential between small and large particles. Considering for example a suspension of particles in a solvent, the equilibrium between precipitate and solute will initially be established when the chemical potentials of the solution, μ_{SOL} , and that of the *smallest* particle, $\mu_{r,small}$, are the same. However, precipitation will still

occur locally at the surface of the larger particles, since $\mu_{r,large} < \mu_{r,small}$ and $\mu_{r,small} = \mu_{SOL}$. This precipitation, however, will further reduce the chemical potential in the solution, so now $\mu_{SOL} < \mu_{r,small}$, leading to further dissolution of the smaller particles. The process continues, and smaller particles dissolve while large particles grow.

One of the most important effects of the coupling between size and chemical potential, especially because of its practical consequences in nanotechnology processes, is the dramatic drop of the melting temperature, T_m , with size. For a spherical particle, mathematical manipulation of the Gibbs-Thomson equation leads to the following relationship:

$$\frac{T_m(r)}{T_{m,\infty}} = \left(1 - \frac{2\gamma\Omega}{H_f r}\right) \quad (16)$$

where r is the particle radius, $T_{m,\infty}$ is the melting temperature of the bulk, H_f is the enthalpy of fusion. As mentioned, this has a major effect also on the sintering processes of nanopowders and in general of nanostructured materials.

The size of a system can also influence *thermodynamic phase stability* at a given temperature. This can again be derived by considering that the energy of the surface increases when reducing size. Consider a bulk material that is thermodynamically stable in a phase A at higher temperature and in a phase B at lower temperature. Considering now a spherical particle of radius r , the free energy change associated to the transformation from phase A to phase B must include surface terms as well as bulk terms:

$$\Delta G_{A \rightarrow B} = G_B - G_A = \left(\frac{4}{3} \pi r^3 \frac{\rho N_A}{M} \mu_B + 4\pi r^2 \gamma_B \right) - \left(\frac{4}{3} \pi r^3 \frac{\rho N_A}{M} \mu_A + 4\pi r^2 \gamma_A \right) \quad (17)$$

Here the factor $(4\pi r^3/3)(\rho N_A/M)$ multiplying the chemical potential is simply the number N of atomic or molecular constituents in the particle, as defined before; M is the molar mass, ρ the density,

and N_A the Avogadro number. First, one can observe that $\mu_B < \mu_A$ because in the bulk, below the transformation temperature, A will spontaneously transform to B. Then, by setting $\Delta G_{A \rightarrow B} = 0$, simple mathematical manipulation leads to the result that, if $\gamma_B > \gamma_A$, there exists a critical radius:

$$r_c = 3 \frac{M}{\rho N_A} \frac{\gamma_B - \gamma_A}{\mu_A - \mu_B} \quad (18)$$

such that if $r < r_c$ then $\Delta G_{A \rightarrow B} > 0$. In other words, particles smaller than the critical radius are thermodynamically stable, and the transformation from A to B does not occur even if the system is below the bulk transformation temperature. Clearly, this effect is only important when the surface terms in Eq. 17 are of the same magnitude as the volume terms, and this can only happen for small radii. Considering typical values of γ and g , r_c is in the order of a few nanometers. A typical example [5] is the stabilization, obtained by reducing the grain size below the critical radius, of tetragonal zirconia at room temperature – where the thermodynamically stable phase for the bulk would be monoclinic. The critical size for pure zirconia powders is in the range of 5–10 nm, while in a polycrystalline zirconia (where stress effects arise) it is about 30 nm. However, for the practical use of zirconia, chemical stabilizers are commonly added, too, such that stabilization is achieved for larger grain sizes, compatible with common processing.

S

Nanothermodynamics

Although quite effective in achieving useful and realistic results, the approach outlined above for studying the thermodynamics of small systems – consisting in adding surface-related ad-hoc terms to the energy equations – has a certain degree of inconsistency: When dealing with macroscopic systems, energy is an extensive quantity, i.e., proportional to the amount of material that constitutes the system. This does not hold anymore in nanoscopic systems, since a large portion of the total energy of the system is

associated to the surface, which is dependent on the size of the objects that constitute the system. Hill has proposed a consistent thermodynamic approach [6, 7], overcoming these inconsistencies.

One can consider a macroscopic system comprising η equivalent and non-interacting, smaller subsystems (a good example would be a colloidal suspension of nanocrystals), characterized by the extensive variables U, S, V, N as defined before. Then, for the entire system $U_{tot} = \eta U; V_{tot} = \eta V; S_{tot} = \eta S, N_{tot} = \eta N$. Hill modifies the standard thermodynamic equation for the entire system as:

$$dU_{tot} = TdS_{tot} - PdV_{tot} + \mu dN_{tot} + Ed\eta \quad (19)$$

Essentially, he adds the term $Ed\eta$, where $E = \partial U_{tot}/\partial\eta|_{S_{tot}, V_{tot}, N_{tot}}$ is the *subdivision potential*. The procedure followed by Hill is analogous to what Gibbs had done by introducing the chemical potential. The analogy is even stronger, however: While the chemical potential determines how the energy of the system varies when changing the number N_{tot} of atoms or molecules that constitute the system, the subdivision potential determines the energy change upon a variation of the number of subsystems η – which in turn corresponds to a variation of the size of the subsystems if the total amount of material in the system is kept constant. In nanoscopic systems, changes in the number of subsystems imply large variations to the total energy, and the subdivision potential contributes significantly to Eq. 19. However, the subdivision potential will be essentially zero for macroscopic systems, as a change in the number of subsystems will have little effect on the total energy (considering for example the surface energy contribution, this will be negligible because in the surface area variations are small as long as the system size is not in the nanoscale regime). In this case, Eq. 19 reduces to the standard thermodynamic equation, so that in this view nanothermodynamics can be viewed as a generalization of standard thermodynamics. Without recurring to ad-hoc additions to the thermodynamic equations, the subdivision potential enables

one to naturally and consistently treat all size-related phenomena, such as surface effects, edge effects, rotation and translation of the systems, etc., thus becoming particularly useful in nanosystems and especially when such phenomena occur together and possibly when coupled. A number of complex nanoscale problems have been successfully treated by using nanothermodynamics, most notably: open linear biological aggregates, local correlations in bulk magnetic materials, glassy systems, and metastable liquid droplets in vapor. More formulations of nanothermodynamics have been formulated over the past couple of decades, with some advantages in terms of the treatment of fluctuations – which might be important in nanosystems – or for the treatment of complex systems, but are essentially equivalent to Hill's.

Cross-References

- Applications of Nanofluidics
- Capillary Flow
- Lotus Effect
- Mechanical Properties of Nanocrystalline Metals
- Nanomechanical Properties of Nanostructures
- Nanoscale Properties of Solid–Liquid Interfaces
- Nanostructures for Surface Functionalization and Surface Properties
- Nanotribology
- Self-Assembled Monolayers for Nanotribology
- Self-Assembly
- Self-Assembly of Nanostructures
- Surface Tension Effects of Nanostructures
- Surface-Modified Microfluidics and Nanofluidics
- Wetting Transitions

References

1. Gibbs, J.W.: Collected Works, vol. 1. Yale University Press, New Haven (1948)
2. Zangwill, A.: Physics at Surfaces. Cambridge University Press, New York (1988)

3. Porter, D.A., Easterling, K.E.: Phase Transformations in Metals and Alloys. Van Nostrand Reinhold, Wokingham (1981)
4. Cao, G.: Nanostructures & Nanomaterials: Synthesis, Properties & Applications. Imperial College Press, London (2004)
5. Garvie, R.C.: Stabilization of the tetragonal structure in zirconia microcrystals. *J. Phys. Chem.* **82**, 218 (1978)
6. Hill, T.L.: A different approach to nanothermodynamics. *Nano Lett.* **1**, 273–275 (2001)
7. Hill, T.L.: Thermodynamics of Small Systems. Dover, New York (1994)

Surface Energy Density

► Surface Tension Effects of Nanostructures

Surface Energy Density and Chemical Potential at Nanoscale

► Surface Energy and Chemical Potential at Nanoscale

Surface Engineering, Tailored Wettability, and Applications

Solomon Adera, Jiansheng Feng and Evelyn N. Wang

Device Research Laboratory, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Synonyms

Contact angle; Hydrophilic; Hydrophobic; Three-phase contact line

Definition

Wettability is a fundamental property of solid surfaces, which plays a role in all aspects of our

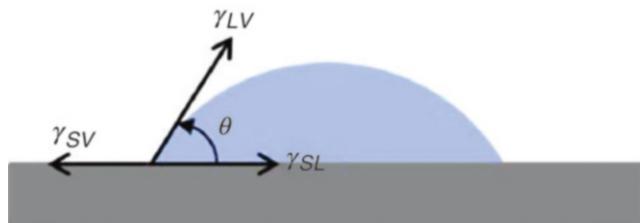
lives. It is characterized by the contact angle that a liquid droplet makes when deposited on a solid surface. The contact angle is the angle subtended by the liquid-vapor and the solid-liquid interface from the liquid side at the three-phase contact line where the three phases (solid, liquid, and vapor) meet. The Young contact angle is the angle that a droplet makes when it comes in contact with an atomically flat, chemically homogeneous, nonreactive, rigid, and insoluble surface (Fig. 1). This angle is typically referred to as the intrinsic or equilibrium or static contact angle (θ) and can be obtained by balancing the interfacial forces at the three-phase contact line as given by the classical Young equation [1]

$$\cos \theta = \frac{\gamma_{SV} - \gamma_{SL}}{\gamma_{LV}}. \quad (1)$$

Here, γ is the interfacial energy per unit area (or equivalently the force per unit length) that acts along the interface indicated by the respective indices (S for solid, L for liquid, and V for vapor). For example, γ_{LV} is the force per unit length along the liquid-vapor interface. Thus, Eq. 1 shows that the equilibrium contact angle for a smooth surface is determined by the chemical nature of the different phases involved. For a water droplet, if this equilibrium contact angle is less than 90° , the surface is termed hydrophilic (wetting), and if the equilibrium contact angle is greater than 90° , the surface is termed hydrophobic (non-wetting).

Introduction

Surface texturing with micro-/nanostructures has been established as an effective way for achieving desirable fluidic (e.g., wetting, self-cleaning) [2], optical (e.g., antireflection) [3], mechanical (e.g., hydrodynamic drag reduction) [4], and thermal (e.g., heat transfer and phase change) [5] properties. These applications, however, require superior wetting characteristics which is a long standing challenge and current research interest in surface engineering. Two extreme limits are often desired: complete wetting and non-wetting behaviors. In the limit of complete wetting, a liquid droplet



Surface Engineering, Tailored Wettability, and Applications, Fig. 1 Schematic of a droplet deposited on a smooth solid surface showing interfacial forces acting

spontaneously spreads with a near zero contact angle forming a thin liquid film. On the other hand, in the limit of complete non-wetting, a liquid droplet remains contained and maintains a spherical shape with high contact angle ($\theta > 150^\circ$) and minimal solid-liquid contact area.

Micro-/nanofabrication is used to improve surface wettability. By introducing roughness, an intrinsically hydrophilic surface can be rendered more hydrophilic by reducing the apparent contact angle, while an intrinsically hydrophobic surface can be rendered more hydrophobic by increasing the apparent contact angle [6]. Consequently, micro-/nanostructuring offers exciting opportunities in tailoring the wetting characteristics of surfaces for various engineering applications ranging from liquid transport to heat transfer, where insight into the fundamental physics and mechanistic understanding of the wetting dynamics are essential, as explained in this entry.

Liquid Spreading Dynamics

Symmetry of the boundary conditions of a liquid droplet deposited on a solid surface dictate that the spreading process as well as the final shape should be axisymmetric. However, chemical heterogeneity and surface roughness, which are typical of actual surfaces, can introduce local energy barriers that can cause contact line pinning resulting in asymmetric propagation dynamics. This asymmetry and liquid propagation on micro-/nanostructured surfaces has been extensively studied due to its applications in microfluidics and lab-on-a-chip devices. However, the liquid transport and wetting dynamics on these surfaces, which are influenced by surface tension forces due

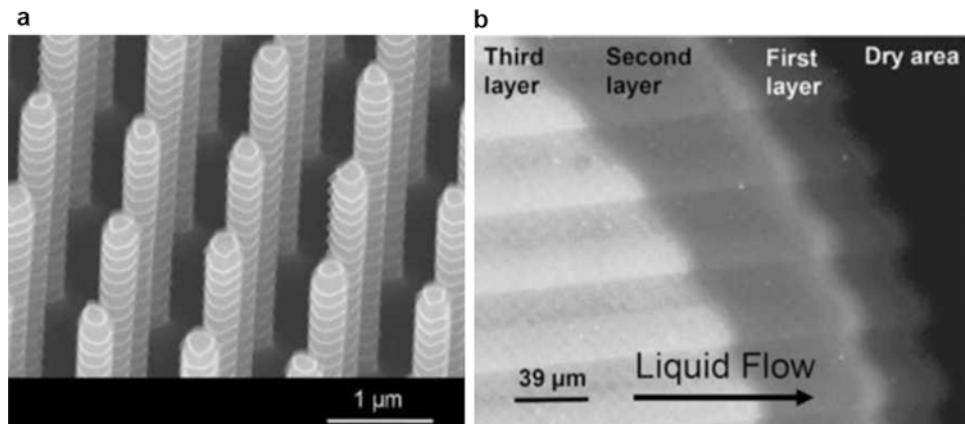
at the three-phase contact line. The equilibrium contact angle is determined by balancing the interfacial forces at the contact line

to the small length scale, are complex and not always well understood leaving room for speculation to the governing physics. In this section, we describe two examples where structure length scale and induced roughness can play important roles in unique liquid spreading behavior.

Multilayer Liquid Spreading in Micropillar Arrays

Recent experimental investigations show that silicon micropillar arrays with scallop features can cause a spreading or receding liquid front to separate into multiple layers with varying thickness [7] as shown in Fig. 2.

The macroscopic layer-by-layer liquid spreading is due to the presence of scallop features on the side walls of the pillars which result from the deep reactive ion etching process that is used to create the micropillars. The scallop features act as energy barriers during the spreading process by pinning the three-phase contact line at the sharp edges on the sides of the pillars. The propagating liquid overcomes these energy barriers in a stepwise manner. When the liquid overcomes the first energy barrier, it forms the first liquid layer that propagates forward and wets the surface. Subsequent layers follow as the liquid climbs up the scalloped tiers by sequentially overcoming the respective energy barriers. This results in subsequent layers of liquid propagating on top of the first layer resulting in multilayer liquid spreading. For pillars with the same diameter and spacing but in the absence of nonvisible scallop features, the liquid spreads uniformly in one layer across the



Surface Engineering, Tailored Wettability, and Applications, Fig. 2 Multilayer liquid spreading in micropillar arrays. (a) Scanning electron micrograph of a surface with nanostructured arrays of pillars with diameter of 500 nm and spacings of 800 nm. The scallops, which are a result of the Bosch deep reactive ion etching process to

entire micropillar array structure with no distinct layers.

This multilayer liquid spreading on silicon micropillar arrays can be explained using a surface energy-based thermodynamic model. The model treats the scallop features as tiered steps that present an energy barrier in the liquid propagation dynamics. The analytical model provides design guideline for creating micropillar arrays that promote multilayer liquid spreading and offers opportunities to control the propagation dynamics as well as the liquid film thickness on superhydrophilic surfaces.

Unidirectional Liquid Spreading

Unidirectional spreading of liquids has been experimentally demonstrated on a macroscopically uniform surface which consists of arrays of asymmetric pillars which are collectively orientated in a single direction [8]. The slanted nanopillars allow the wetting liquid to propagate in one direction only (Fig. 3a, b), i.e., along the pillar-tilting direction (denoted as +X, Fig. 3c). In the opposite direction ($-X$), the three-phase contact line is pinned at the base of the nanopillars, and the liquid does not spread. To make such a

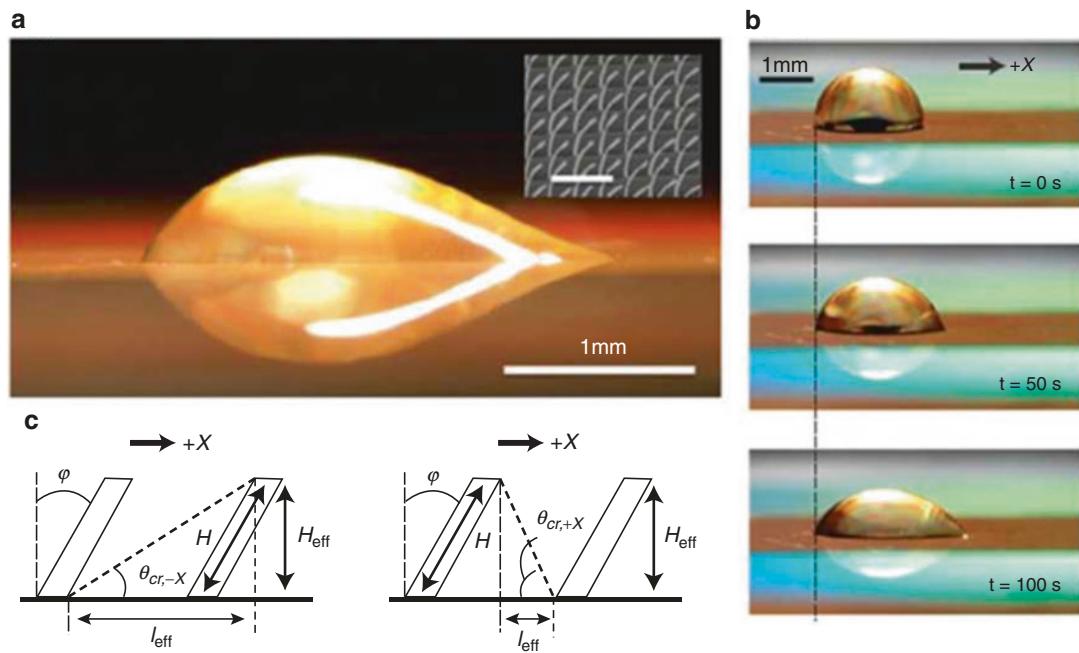
create deep trenches, have dimensions of approximately 100 nm. (b) Visualization of multilayer spreading on the corresponding geometry in (a). Liquid spreads from *left* to *right* where the dark area is dry. The differences in brightness indicate variations in liquid film thickness (Reproduced with permission from Xiao et al. [7])

surface, first arrays of upright silicon nanopillars were fabricated using deep reactive ion etching. This is followed by depositing a thin gold film on one side of the nanopillars using electron-beam evaporation. When cooled to room temperature, the residual thermal stress on the thin gold film caused the partially coated silicon nanopillars to deflect in one direction. The deflection angle is observed to be a function of the thickness of the gold film.

When a wetting droplet is deposited on these surfaces, the slanted arrays of nanopillars pull the liquid toward the pillar top. As the meniscus at the spreading front is pulled forward by one row of pillars, the contact line at the base of the substrate reaches the next row of pillars. Modeling results show that the liquid continues to spread if the intrinsic contact angle is below a critical angle that is determined from the pillar array geometry. In the opposite direction, however, the contact line is pinned at the base of the front row of nanopillars, and the liquid does not spread resulting in unidirectional spreading.

Wettability and Heat Transfer

In boiling heat transfer, textured hydrophilic surfaces have been demonstrated to enhance the thermal performance. Similarly, condensation heat



Surface Engineering, Tailored Wettability, and Applications, Fig. 3 Unidirectional liquid spreading on slanted pillar array. (a) The characteristics of a spreading droplet on asymmetric nanostructured surface (*inset*) at one instant in time. The scale bar in the *inset* is 10 μm . The diameter, spacing, height, and deflection angle of the nanopillars are 0.5 μm , 3.5 μm , 10 μm , and 12°, respectively. (b) Side view time-lapse images of unidirectional

spreading of a 1 μL droplet of deionized water with 0.002 % by volume of surfactants (Triton X-100). The initial location of the droplet contact line in the $-X$ direction is indicated by the *dashed line*. (c) Schematic diagrams explaining the geometries for the model to determine the critical angle in the $-X$ (*left*) and $+X$ (*right*) directions (Reproduced with permission from Chu et al. [8])

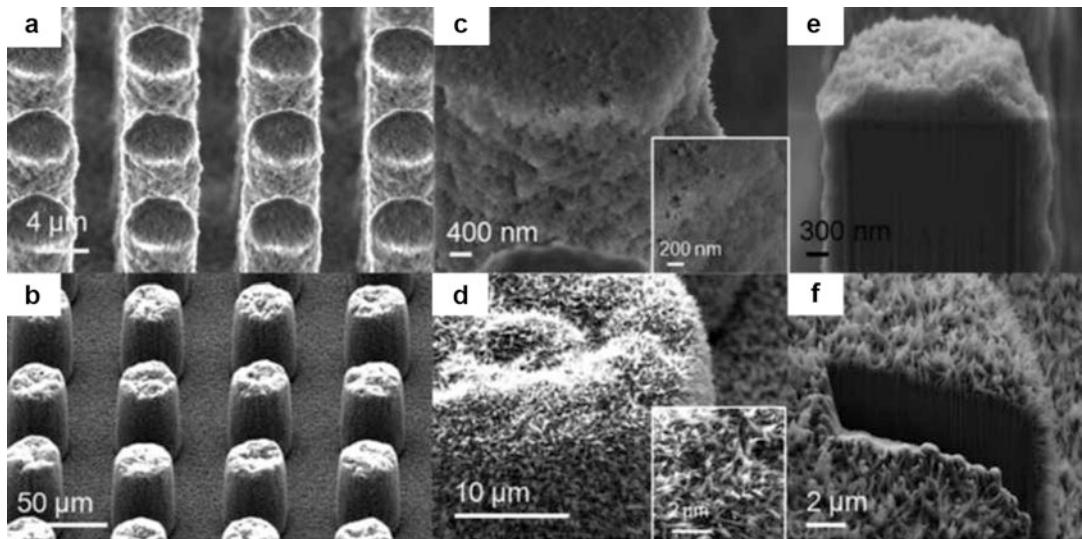
transfer has been improved by using textured hydrophobic surfaces. The enhancements in both cases are attributed to the induced roughness and improved wettability. In this section, we briefly discuss three examples where surface wettability affects the heat transfer and phase-change process, sometimes even in unexpected ways.

Hierarchical Surfaces for Pool Boiling Heat Transfer

Boiling heat transfer is used in various large-scale industrial applications including water purification, desalination, and power generation. Consequently, modest enhancements in boiling heat transfer would lower energy consumption and cost by making the system more efficient.

Boiling heat transfer is driven by microlayer and interline evaporation, transient conduction, and microconvection as bubbles nucleate, grow, and depart from the heated surface [9]. As the heat flux increases, however, a competition arises between the vapor generated at the nucleation site and the replenishing of the liquid. At a certain point, the system reaches an operational limit beyond which the boiling heat transfer performance decreases due to the formation of a thermally insulating vapor layer that separates the heated surface from the working fluid [10]. This operational limit is termed the critical heat flux (CHF), and it signifies the beginning of transition boiling.

Recent advances in micro/nanofabrication techniques have produced high-surface-area superhydrophilic surfaces that are capable of



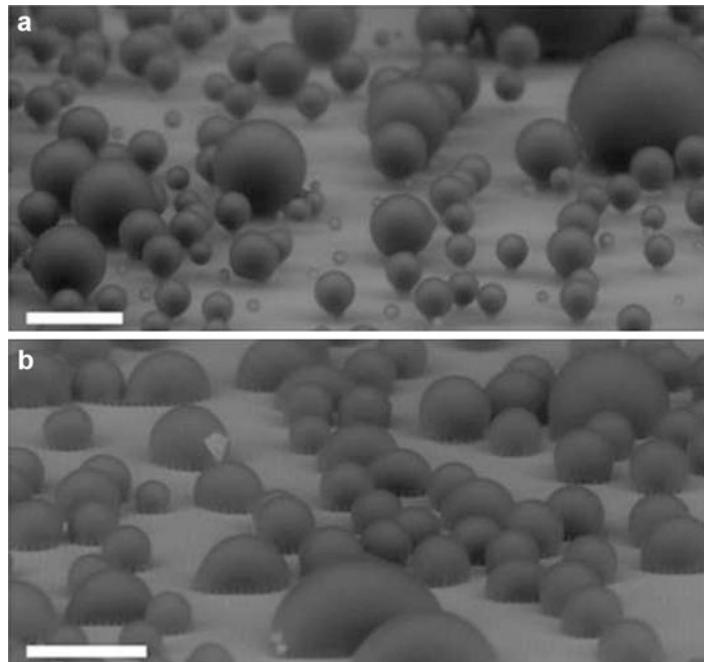
Surface Engineering, Tailored Wettability, and Applications, Fig. 4 Scanning electron micrograph of the representative, fabricated silica- and CuO-based hierarchical surfaces. (a) Electrophoretic deposition (EPD)-coated silica; micropillars in a square array with heights of 10 μm , diameters of 10 μm , and spacings of 5 μm ; and (b) CuO micropillars in a square array with heights of 61 μm , diameters of 30 μm , and spacings of 30 μm .

(c) Magnified view of the silica-based micropillar and EPD-coated SiO_2 nanoparticles (inset). (d) Magnified view of the CuO micropillar and CuO nanostructures formed on the surfaces (inset). (e) Cross section view of the silica-based micropillar and (f) cross section view of the CuO micropillar obtained using focused ion beam milling (Reproduced with permission from Chu et al. [12])

increasing the CHF limit [11, 12]. Single length scale microstructured surfaces have increased the CHF of water to $\approx 150\text{--}210 \text{ W/cm}^2$. Chu et al. [12] further increased the CHF to $\approx 250 \text{ W/cm}^2$ using multiple length scale roughness or hierarchical surfaces which were fabricated using electroplated copper microstructures covered with copper oxide (CuO). The CuO hierarchical surfaces have a roughness factor, defined by the ratio of the overall area to the projected area, as high as 13.3. The large enhancements in CHF on the hierarchical surfaces (up to 200 % compared to smooth SiO_2 surfaces) were attributed to the increased roughness factor which provides an even larger surface force to counteract the momentum force from evaporation. Both experiments and modeling show that the roughness-amplified surface forces contribute significantly toward improving the CHF limit in pool boiling on micro-/nanostructured hierarchical surfaces (Fig. 4).

Superhydrophobic Surfaces for Enhanced Condensation

During condensation heat transfer, water droplets preferentially condense on rough solid surfaces (heterogeneous condensation) rather than directly from the vapor phase (homogeneous condensation) due to the smaller energy barrier or activation energy. Condensing vapor on typical industrial surfaces such as copper and stainless steel forms a thin liquid film because of the high surface energy of these metals. As a result, the condensing liquid film creates additional thermal resistance in the heat conduction path lowering the thermal performance and efficiency. This mode of condensation is called filmwise [13]. On the other hand, if the surface is chemically coated to lower the surface energy, the vapor condenses on the surface with discrete liquid droplets. This mode of condensation is called dropwise [14]. In dropwise condensation, gravity



Surface Engineering, Tailored Wettability, and Applications, Fig. 5 Droplet morphology during condensation. Condensed droplet growth observed using environmental scanning electron microscope (ESEM) on Si nanopillars with (a) Cassie droplets with pitch spacing $l = 2 \mu\text{m}$ and (b) Wenzel droplets with $l = 4 \mu\text{m}$. The

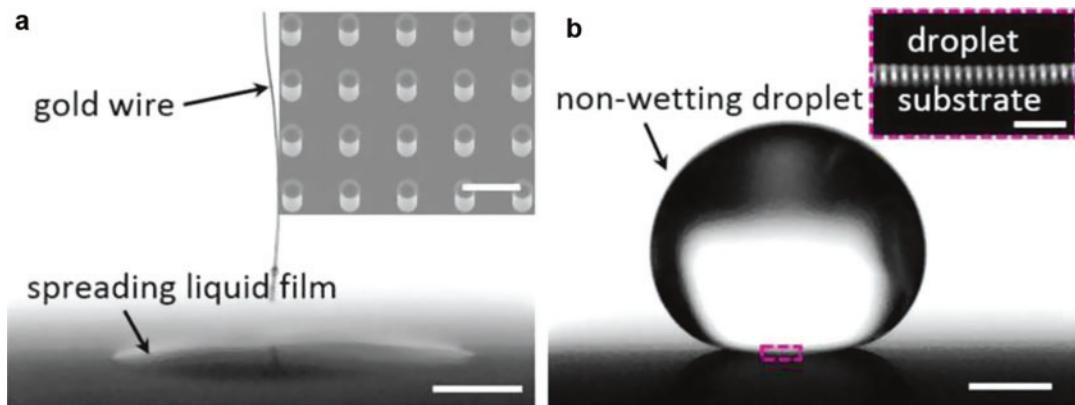
diameter and height of the pillars were $d = 300 \text{ nm}$ and $h = 6.1 \mu\text{m}$, respectively. The surfaces were functionalized with a silane having an intrinsic advancing contact angle of $\theta_a \approx 22^\circ$. Scale bars = $60 \mu\text{m}$ (Reproduced with permission from Enright et al. [18])

removes the condensing droplets when they reach the capillary length ($\approx 2.7 \text{ mm}$ for water) and refreshes the surface preparing it for re-nucleation. As a result, dropwise condensation has been demonstrated to improve the thermal performance by five to seven times when compared to filmwise condensation [15].

Recent advances in micro-/nanofabrication have enabled the development of micro-/nanostructured hydrophobic surfaces that allow greater mobility of condensing droplets (Fig. 5). If properly designed, the small length scale and solid fraction of these superhydrophobic surfaces enable the condensing droplets to move without significant resistance, promoting the merging of neighboring droplets. When droplets merge and form a bigger droplet, the excess surface energy created during coalescence is converted to kinetic energy enabling the merged droplet to jump off the surface. This mode of condensation is called jumping droplet condensation [16]. The

coalescence-induced droplet ejection can greatly improve the overall efficiency of the condensation process on superhydrophobic surfaces [16]. Compared to state-of-the-art dropwise condensation on copper surfaces, jumping droplet condensation has demonstrated a 30 % increase in the heat transfer coefficient ($\approx 92 \text{ kW/m}^2\text{-K}$) [17]. The initiation and growth of the condensate droplet depends on the surface-droplet and droplet-droplet interaction [18]. A unified model coupling individual droplet heat transfer, droplet size distribution, and wetting morphology suggests that a range of geometries from 0.5 to $2 \mu\text{m}$ promise a 190 % overall heat flux enhancement of jumping droplet condensation over conventional dropwise condensation on flat surfaces [17].

Jumping droplet condensation, however, cannot be sustained when the nucleation density is very high compared to structure density. In such cases, neighboring droplets coalesce and form discrete non-jumping droplets that strongly



Surface Engineering, Tailored Wettability, and Applications, Fig. 6 (a) A droplet deposited on a superhydrophilic surface at 120 °C spontaneously spreads into a thin film and wets the surface. The *inset* shows the scanning electron micrograph of the microstructured surface. Scale bars in the figure and the *inset* are 0.5 mm and 20 μ m, respectively. (b) A similar-sized droplet at an

elevated temperature (\approx 160 °C) did not wet the same surface; instead, it rested on top of the microstructured surface. The *inset* shows a magnified view of the boxed section near the droplet base indicating that the droplet remained in contact with the pillar tops. Scale bars in the figure and the *inset* are 0.5 mm and 100 μ m, respectively (Reproduced with permission from Adera et al. [20])

adhere to the surface. This mode of condensation is called flooding [17]. The flooding creates an additional thermal resistance to the heat conduction path that degrades the thermal performance and reduces the heat transfer coefficient (\approx 44 kW/m²-K) [17].

Evaporation-induced Non-wetting Water Droplets on Superhydrophilic Surfaces

Microstructuring along with chemical functionalization is extensively used in enhancing surface wettability at room temperature [6, 19]. However, the wetting behavior of droplets at room temperature does not necessarily apply to microstructured surfaces heated above the saturation temperature. Unlike their wetting behavior at room temperature, water droplets exhibit drastically different and yet potentially useful non-wetting behavior on nominally hot superhydrophilic surfaces [20].

A water droplet deposited on a rough hydrophilic surface at room temperature spreads spontaneously with vanishing contact angle due to the induced roughness (Fig. 6a). However, when the microstructured hydrophilic (i.e.,

superhydrophilic) surface is heated slightly above the saturation temperature, it can sustain non-wetting water droplets with contact angles as high as \approx 160° (a superhydrophobic behavior) due to induced evaporation (Fig. 6b) [20]. The temperature at which this change in the wetting behavior occurs is significantly lower than the classical Leidenfrost temperature [10]. Moreover, the evaporation-induced non-wetting water droplets remain in contact with the pillar tops (inset Fig. 6b) making them distinctively different from Leidenfrost drops where solid-liquid contact is absent due to the presence of a thin (10–100 μ m) [21] and yet stable vapor cushion.

The evaporation-induced non-wetting behavior of droplets on hot superhydrophilic surfaces is explained by using a model that analyzes the forces acting on the suspended droplet. Surface tension force acts to induce wetting due to the hydrophilicity of the surface. The heat conduction through the pillars causes the droplet to evaporate from its base. Due to the low permeability of the porous media, the radially escaping vapor underneath the droplet gives rise to a pressure gradient and an upward pushing force that counteracts wetting. By balancing the two competing forces (surface tension and pressure), the model predicts the superheat that is required for a water droplet to

reside on an arrays of silicon micropillars without wetting.

The presence of micropillar arrays has increased the effective thermal conductivity and decreased the vapor permeability of the porous media. As a result, water droplets deposited on nominally heated microstructured hydrophilic surfaces remain non-wetting. The mechanistic insight gained from this study provides design guidelines. The use of low thermal conductivity, high aspect ratio, and small-scale, highly porous structures is useful for applications requiring wetting droplets. Conversely, large-scale structures with high thermal conductivity should be used to reduce hydrodynamic drag on nominally heated microstructured hydrophilic surfaces.

Cross-References

► Wetting Transitions

References

1. Young, T.: An essay on the cohesion of fluids. *Philos. Trans. R. Soc. Lond.* **95**, 65–87 (1805)
2. Seemann, R., Brinkmann, M., Kramer, E.J., Lange, F.F., Lipowsky, R.: Wetting morphologies at microstructured surfaces. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 1848–1852 (2005)
3. Huang, Y.-F., et al.: Improved broadband and quasi-omnidirectional anti-reflection properties with biomimetic silicon nanostructures. *Nat. Nanotechnol.* **2**, 770–774 (2007)
4. Vakarelski, I.U., Patankar, N.A., Marston, J.O., Chan, D.Y., Thoroddsen, S.T.: Stabilization of Leidenfrost vapour layer by textured superhydrophobic surfaces. *Nature* **489**, 274–277 (2012)
5. Launay, S., Fedorov, A., Joshi, Y., Cao, A., Ajayan, P.: Hybrid micro-nano structured thermal interfaces for pool boiling heat transfer enhancement. *Microelectron. J.* **37**, 1158–1164 (2006)
6. Quéré, D.: Wetting and roughness. *Ann. Rev. Mater. Res.* **38**, 71–99 (2008)
7. Xiao, R., Chu, K.-H., Wang, E.N.: Multilayer liquid spreading on superhydrophilic nanostructured surfaces. *Appl. Phys. Lett.* **94**, 193104 (2009)
8. Chu, K.-H., Xiao, R., Wang, E.N.: Uni-directional liquid spreading on asymmetric nanostructured surfaces. *Nat. Mater.* **9**, 413–417 (2010)
9. Kandlikar, S.G.: A theoretical model to predict pool boiling CHF incorporating effects of contact angle and orientation. *J. Heat Transf.* **123**, 1071–1079 (2001)
10. Leidenfrost, J.G.: On the fixation of water in diverse fire. *Int. J. Heat Mass Transf.* **9**, 1153–1166 (1966)
11. Chu, K.-H., Enright, R., Wang, E.N.: Structured surfaces for enhanced pool boiling heat transfer. *Appl. Phys. Lett.* **100**, 241603 (2012)
12. Chu, K.-H., Joung, Y.S., Enright, R., Buie, C.R., Wang, E.N.: Hierarchically structured surfaces for boiling critical heat flux enhancement. *Appl. Phys. Lett.* **102**, 151602 (2013)
13. Kutateladze, S., Gogonin, I.: Heat transfer in film condensation of slowly moving vapour. *Int. J. Heat Mass Transf.* **22**, 1593–1599 (1979)
14. Bonner, R.W.: 2010 14th International Heat Transfer Conference 221–226 (American Society of Mechanical Engineers)
15. Sikarwar, B.S., Khandekar, S., Agrawal, S., Kumar, S., Muralidhar, K.: Dropwise condensation studies on multiple scales. *Heat Trans. Eng.* **33**, 301–341 (2012)
16. Boreyko, J.B., Chen, C.-H.: Self-propelled dropwise condensate on superhydrophobic surfaces. *Phys. Rev. Lett.* **103**, 184501 (2009)
17. Miljkovic, N., Wang, E.N.: Condensation heat transfer on superhydrophobic surfaces. *MRS Bull.* **38**, 397–406 (2013)
18. Enright, R., Miljkovic, N., Al-Obeidi, A., Thompson, C.V., Wang, E.N.: Condensation on superhydrophobic surfaces: The role of local energy barriers and structure length scale. *Langmuir* **28**, 14424–14432 (2012)
19. Wenzel, R.N.: Resistance of solid surfaces to wetting by water. *Ind. Eng. Chem.* **28**, 988–994 (1936)
20. Adera, S., Raj, R., Enright, R., Wang, E.N.: Non-wetting droplets on hot superhydrophilic surfaces. *Nat. Commun.* **4**, 2518 (2013)
21. Biance, A.-L., Clanet, C., Quere, D.: Leidenfrost drops. *Phys. Fluids* **15**, 1632–1637 (2003)

Surface Figuring

► Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Surface Force Balance

► Surface Forces Apparatus

Surface Forces Apparatus

Carlos Drummond¹ and Marina Ruths²

¹Centre de Recherche Paul Pascal,
CNRS–Université Bordeaux 1, Pessac, France

²Department of Chemistry, University of
Massachusetts Lowell, Lowell, MA, USA

Synonyms

Surface force balance

Definition

The surface forces apparatus (SFA) is an instrument for sensitive measurements of normal and lateral forces between two macroscopic surfaces in contact or separated by a thin film. The surface separation distance can be measured and independently controlled to 0.1 nm. The surfaces typically form a single-asperity contact where the substrates deform elastically, and time- and rate-dependent effects in the measured forces can be ascribed to phenomena in the thin films or adsorbed monolayers confined between the surfaces.

Basics of the SFA Technique

The Surfaces and the Apparatus

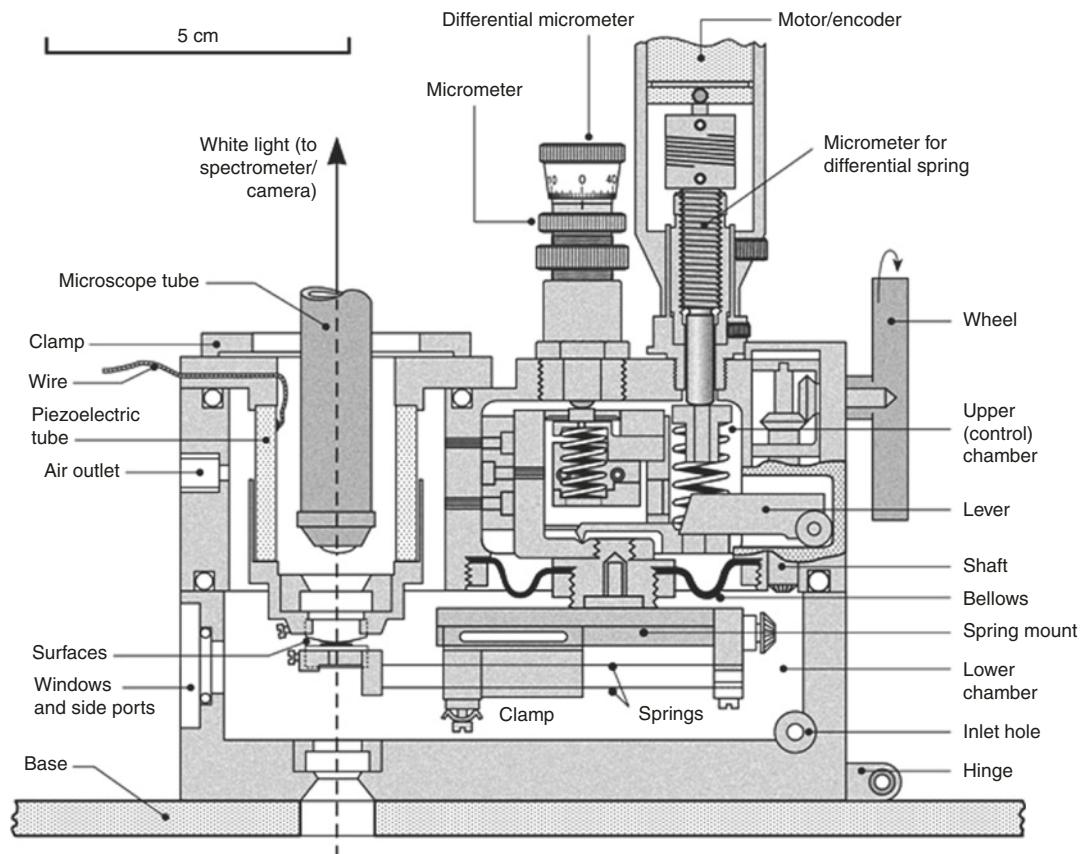
In SFA experiments, normal and lateral interaction forces are measured between two surfaces across air or a medium (a confined film). The most commonly used surface substrates are back-silvered, molecularly smooth muscovite mica sheets glued to half-cylindrical fused-silica disks. These half-cylindrical surfaces are mounted in the SFA in a crossed-cylinder configuration (cf. Figs. 1 and 2). At surface separations (distances D) much smaller than the radius of curvature (R) of the surfaces, this is equivalent to

a sphere-on-flat geometry, which presents several advantages over that of two parallel plates: Alignment issues and edge effects are avoided, and different points of contact between the two surfaces can be easily investigated by displacing the disks laterally (e.g., to check the repeatability of measurements at different regions of a pair of surfaces or to move away from wear debris or contamination). In addition, the normal force F between cylindrical surfaces is related to the interaction energy between flat surfaces W according to the Derjaguin approximation [4], $F(D)/R = 2\pi W$, which provides a convenient means for comparison of experimental data with model predictions. For this reason, and to allow quantitative comparison of data from different experiments, normal force data are typically presented normalized by R .

Several different SFA setups have been developed for the measurements of normal and lateral forces. Early designs by Tabor, Winterton, and Israelachvili [2] were refined for measurements in liquids and vapors (the Mk II model) [4, 5]. More recent models are easier to assemble and clean (the Mk IV [6]) or have improved distance controls and many different attachments (the Mk III/SFA3 [1] in Figs. 1 and 2 and the SFA2000 [7]). The newer designs are also more user-friendly and have improved stability against thermal and mechanical drifts.

Normal Distance and Force Measurements

The interaction forces normal to the surfaces are determined from the deflection of a spring supporting the lower surface (Fig. 1). Typically, a double-cantilever spring is used to minimize tilting and/or sliding of the surfaces. Its spring constant, k , is determined by putting small weights on its end and measuring the resulting deflection with a traveling optical microscope. During an experiment to determine normal force, F , as a function of distance (surface separation), D , the spring deflection is found by monitoring the change in distance with multiple beam interferometry (MBI) [8] as the surfaces are moved



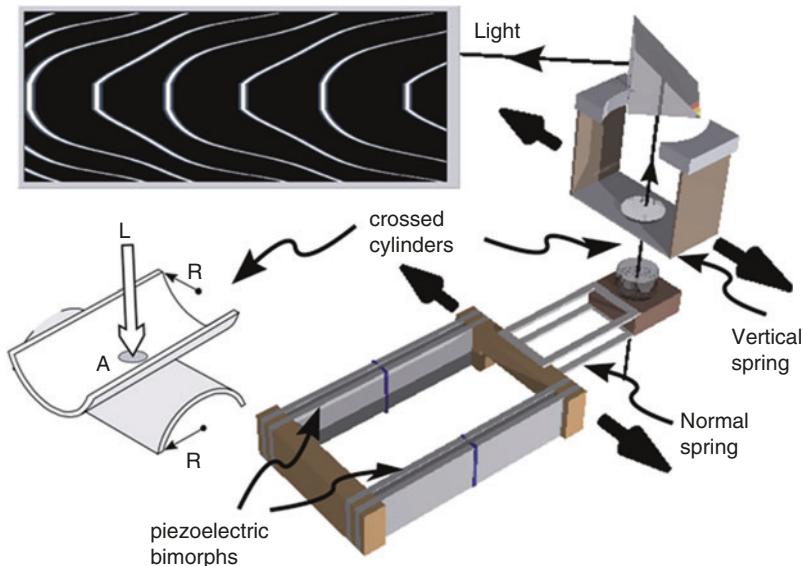
Surface Forces Apparatus, Fig. 1 Schematic drawing of the SFA3 (Mk III) configured for measurements of forces normal to the surfaces [1]. The lower surface is supported on a “force-measuring” double-cantilever spring with movable clamping, and the upper surface is mounted on a holder containing a piezoelectric tube for fine control of the separation distance between the surfaces, D . The

mechanical distance control system (a system of weak and stiff springs reducing the movement of the manual and motor-driven micrometers) is separated from the main chamber of the instrument by Teflon bellows. The path of the light used for multiple beam interferometry (MBI) is indicated with an arrow (Reprinted with permission from Ref. [2])

toward or away from one another using motorized stages (for moving the base of the spring, the “spring mount” in Fig. 1) or a piezoelectric actuator (for moving the top surface). Any deviation in measured distance from the one expected from a calibration of the movement at large separations (where no forces act between the surfaces) represents a deflection of the spring, ΔD . The change in force when moving from one position to another is $\Delta F = k\Delta D$, according to Hooke’s law.

Regions of a force versus distance curve where the gradient of the force, dF/dD , exceeds k are inaccessible to the technique due to mechanical instability, and the surfaces will spontaneously

jump from one stable region to the next. The choice of spring stiffness is therefore of importance for the detection of different regions of the force curve, and several SFA designs incorporate springs whose length can be changed during the experiment (by moving a clamp along the spring, cf. Fig. 1). The use of force feedback to control the force applied to the surfaces independently of the displacement has also been suggested, to maintain a constant deflection of the spring. Implementations of this include the use of a magnetic force transducer and a bimorph deflection sensor [9] or a feedback system utilizing capacitive displacement transducers [10].



Surface Forces Apparatus, Fig. 2 Schematic illustration of the SFA tribometer design by Israelachvili and coworkers. As in Fig. 1, two back-silvered mica surfaces with a radius of curvature R are mounted in a crossed-cylinder configuration. The FECO arising from flattening of these surfaces over a contact area A at a given load L are shown in the top graph. (The alternating shapes of the fringes arise from odd and even orders of interference.) The lower surface is mounted on a double-cantilever

“normal spring” (spring constant k) attached to a piezoelectric bimorph device allowing lateral movement of the lower surface in the direction of the arrows. The upper surface is mounted on a friction sensing device with vertical springs whose deflection is monitored with resistance strain gauges. The direction of incoming and emerging light is indicated with arrows (Adapted with permission from Ref. [3])

Multiple Beam Interferometry (MBI)

The use of multiple beam interferometry (MBI) [8, 11] makes it possible to determine the distance D between the surfaces with subnanometer resolution (0.1 nm) and to measure the refractive index of confined films to 0.01 . MBI is a particular useful feature of the standard SFA setup, since it allows real-time, *in situ* monitoring of the geometry of the interacting surfaces and of the region of contact. Parameters such as the film thickness (including the thickness of individual molecular layers) and the diameter or radius of the contact area (to a lateral resolution of about $1 \mu\text{m}$) can be readily measured. The occurrence of protrusions or regions of different refractive index can also be detected, which may indicate wear or accumulation of material in some region of the contact.

The mica surfaces, any material confined between them, and the semitransparent, highly reflective silver layers on their back sides form a built-in Fabry-Perot interferometer. White light is

passed through the lower surface, and the wavelengths that interfere constructively after multiple reflections between the silver mirrors emerge through the top surface. This light is focused onto the slit of a grating spectrometer. The resulting pattern, fringes of equal chromatic order (FECO, cf. top left of Fig. 2), represents the shape and thickness of the space between the silver layers and depends on the optical path length through the materials in the interferometer (the thicknesses and refractive indexes of the mica sheets and any air gap or material confined between them). A measurement of the interfering wavelengths from surfaces in mica-mica contact in dry N_2 gas is used as the basis for calculation of distances (D) between the front sides of the mica surfaces (and thus of film thicknesses of confined materials), since an increase in optical path causes the fringes to shift to larger wavelengths with respect to this value. Relationships between the measured wavelengths, refractive indexes, and

D for a symmetric interferometer (two mica sheets of equal thickness) [8] have later been extended to more complicated arrangements such as asymmetric, absorbing, anisotropic, or more complicated multilayer systems [11].

SFA-Based Tribometers

Control of Film Thickness and Contact Area

Since the contact geometry in the SFA closely resembles a sphere-on-flat contact and the surfaces can be very smooth (atomically smooth in the case of bare mica surfaces), a single-asperity contact can be formed whose size, position, and deformation are easily monitored using MBI. This contact geometry is also advantageous for comparisons with contact mechanics models [4]. In nanotribological experiments, the film thickness and applied load (normal force) are typically regulated with the motor-driven distance controls. If the surfaces adhere or if the load is sufficiently high, the glue layers under the mica sheets deform elastically [2, 4] and a flat, circular contact area of uniform film thickness T and area of contact A is formed, as shown schematically in Fig. 2 (drawing on left). The corresponding FECO (Fig. 2, top left) show a characteristic flattening indicating a region of uniform film thickness. The diameter of the contact area can be measured directly (to about 1 μm) from this flattened region of the FECO (observations can be made in two orthogonal directions by rotating the light directed to the spectrometer with a dove prism, not shown here). Surfaces with a radius of curvature of $R \approx 2$ cm typically give rise to contacts with a diameter of a few μm to a few hundred micrometers, depending on the load and strength of adhesion. The resulting maximum pressure in the middle of the contact is typically at most a few tens of megapascals.

Any changes in shape of the surfaces or in thickness of the confined film can be observed directly on the FECO and measured with the same resolution as during a measurement of the normal force. Thickening and thinning of the confined film can be observed, as well as shear-induced hydrodynamic deformations.

Furthermore, any damage of the surfaces can be easily detected as soon as it occurs, which enables one to identify sliding conditions that lead to wear and investigate wearless friction versus friction with wear.

Identification and accurate measurements of the contact area are very important for the analysis of nanotribological experiments. In many cases, the measured friction force mainly arises from the region of highest confinement, i.e., from the film in the flattened contact. Certain systems, especially ones where there is adhesion between the surfaces, give rise to a friction force that increases nonlinearly with load and appears to be proportional to this contact area [2]. Unlike many nonadhesive systems, the friction of adhesive contacts is thus not well described by the commonly used friction coefficient, μ (the slope of the friction force as a function of load). They are better characterized by their shear stress, σ , which is the friction force normalized by the contact area. In SFA experiments on a variety of thin films, the shear stress depended on the film structure and thickness and typically remained constant over the investigated range of loads. However, there are situations where the shear stress may vary with pressure, and it is important to recognize that because of the curved surfaces in the SFA, the pressure in the direction normal to the surfaces is not a constant over the flattened contact area but varies in a manner that can be described by contact mechanics models [4].

One can also envision a more complicated situation where regions of the surfaces outside the flattened area contribute to the friction response. This may occur, for example, if molecules are able to bridge a relatively large gap between the surfaces, and their “bonds” to the surfaces have to be broken and reformed during sliding. One possibility to define the contact area is to adopt a cutoff length and assume that the contributions to the frictional force are negligible over larger separation distances. Information on the extension of molecules from the surface can be obtained from measurements of the normal force, but it is still difficult to estimate how a region outside the contact will contribute to the total measured friction force. However, this problem

is shared by most experimental techniques in nanotribology, and the SFA coupled with the MBI technique provides the most direct way to directly observe the contact geometry during sliding.

Shear and Friction Attachments and Measurements

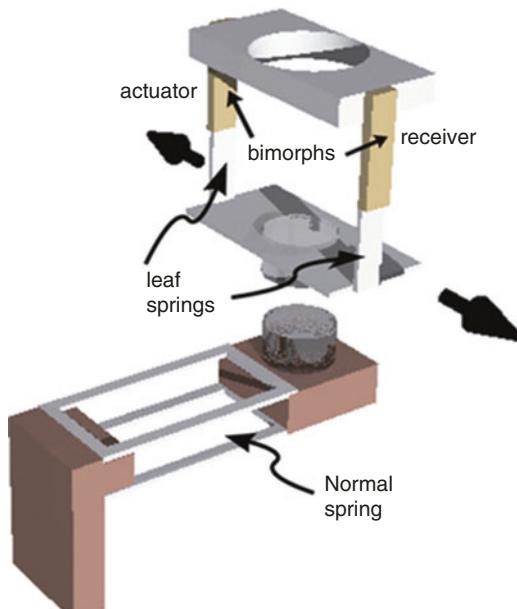
Several types of shear and friction attachments for the SFA have been developed in the past 30 years, each with its own capabilities and limitations. These setups can be used to investigate different regimes of sliding velocity, magnitude of the friction forces, and sliding distance (amplitude). The most commonly used setups are discussed below. All of these setups have in common that the mechanical properties of the system (i.e., its compliance and inertial mass) will influence the results, and these factors have to be taken into account in order to obtain meaningful information from the measured signals. Because of the mechanical simplicity and easily characterized mechanical properties of the SFA, this can be done in a more straightforward manner than in many other devices used for nanotribological investigations. Detailed descriptions of the devices and of experimental data obtained on different systems can be found in the original publications.

Several devices for lateral shear and friction measurements have been developed by Israelachvili and coworkers [3, 7]. A schematic illustration of a current version is shown in Fig. 2. The lower surface is mounted on a double-cantilever spring (of known spring constant k) used to apply and measure the load (normal force) based on expected (calibrated) versus directly measured distance changes. This spring is attached to a “bimorph slider” consisting of sectored piezoelectric elements (electromechanical transducers) [3]. As a constant slope voltage ramp (triangular waveform) is applied over the electrodes of certain sectors of the bimorph elements, the lower surface is translated laterally in a linear manner. The maximum distance (amplitude) of the sliding motion is typically tens of micrometers, depending on the characteristics of the bimorph elements. The driving

speed (sliding velocity of the lower surface) can be varied between a few Å/s to about 0.1 mm/s. The bimorph slider can also be used for nanorheological measurements if a constant frequency sinusoidal input is chosen instead of a triangular wave.

The detection of friction forces or viscoelastic responses is done with the device holding the upper surface [3, 7]. This “friction device” consists of a vertical double-cantilever metal spring (with a known spring constant K), whose deflection is measured using strain gauges forming the arms of a Wheatstone bridge. The force experienced by the upper surface due to the movement of the lower can thus be calculated from this spring deflection using Hooke’s law. The detection limit depends on the stiffness of the spring and sensitivity of the strain gauges and is typically a few μN . In cases where larger displacements are desired than the ones available with the bimorph slider, the friction device can be used as both an actuator and detector: The base of the vertical double cantilever can be translated laterally (approx. ± 5 mm) using a reversible, variable speed motor-driven micrometer (not shown), and the deflection of the vertical springs recorded simultaneously. A combination of the bimorph slider and the friction device, or the friction device alone, has been used by Israelachvili and coworkers to study the shear or friction response of a wide range of systems such as highly confined simple liquids, polymer melts and solutions, self-assembled surfactant and polymer layers, and liquid crystals (see Refs. [2, 4]).

A different design, used to investigate smaller deformations and mainly the linear response of confined films, has been developed by Granick and coworkers [12]. Small deformations are advantageous for investigations of longtime relaxation processes that might be occurring in the contact region. Although mainly developed for the study of deformations of the order of the film thickness or less, the displacement range goes from less than 1 nm to a few hundred nanometers, with a reported force sensitivity of about 5 μN . This type of device is illustrated schematically in Fig. 3. The load (normal force) is regulated by adjusting the vertical position of the lower



Surface Forces Apparatus, Fig. 3 Schematic illustration of the SFA tribometer design by Granick and coworkers. Bending of one of the bimorph elements (the actuator) causes the upper surface to move laterally with a displacement detected with the other bimorph element (the receiver). The distance (film thickness) and load are controlled by moving the base of the spring holding the lower surface, as in the design in Fig. 1 (Adapted with permission from Ref [12])

surface, and this surface remains stationary in the lateral direction during shear experiments.

The upper surface is mounted on a double-cantilever device. Different from the design in Fig. 2, the double cantilever consists not of metal springs with strain gauges but of metal springs with attached piezoelectric bimorph elements. One element is used as an actuator and the other as a detector (receiver). In a typical experiment, a constant frequency sinusoidal input causes bending of the actuator bimorph. The output voltage from the detector bimorph is used to measure the actual displacement of the upper surface. The viscoelastic properties of the confined film is extracted by comparing these signals to those measured with the surfaces in strongly adhesive (mica-mica) contact and when separated, i.e., when the device is moving freely. The electromechanical characteristics of the system are modeled as a combination of effective masses, springs, and

dashpots representing the different components of the apparatus [12]. In practice, the response in mica-mica contact (from the mica and glue layers) is typically purely elastic. This setup has been used by Granick and coworkers to study a variety of systems, including simple liquids, polymer melts, adsorbed layers, and solutions.

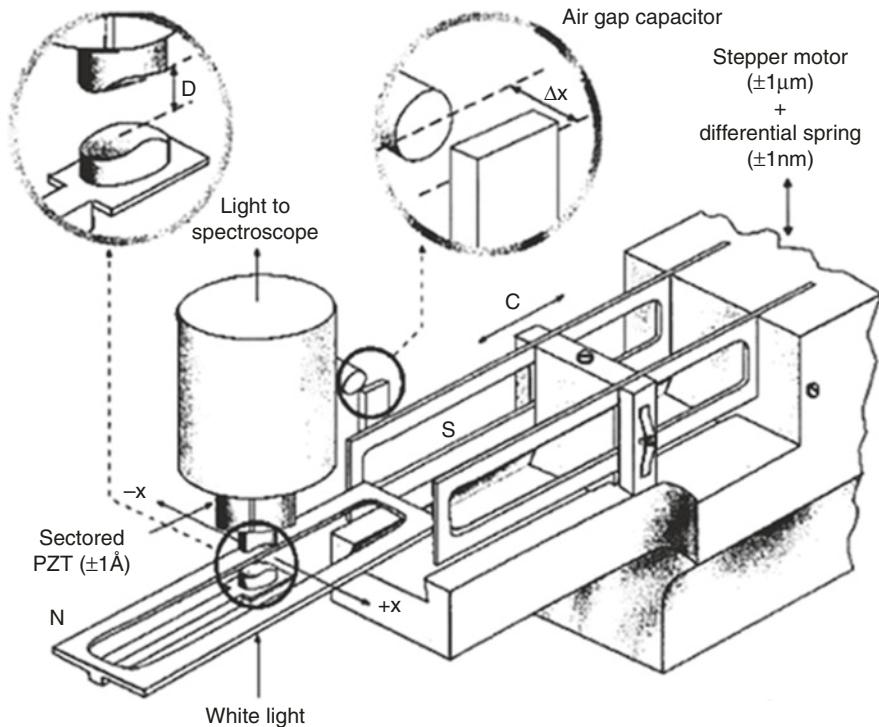
A third type of shear device, developed by Klein and coworkers [13], is shown schematically in Fig. 4. In this design, the displacement is induced by the upper surface, and the friction force is measured at the lower. The upper surface is attached to a sectored piezoelectric tube that is used to produce a normal or lateral displacement. The lower surface is mounted on a single-cantilever spring (for measurements of normal forces, N in Fig. 4), attached to a double-cantilever shear force spring (S in Fig. 4). The lateral displacement of this double-cantilever spring is measured using an air-gap capacitor, and the friction force calculated using Hooke's law. The reported sensitivity of this device is 50 nN, and the maximum displacement of the upper surface is a few tens of micrometers. In this device, the sensitivity to the measured friction forces is greatly improved with respect to the designs discussed above. The high sensitivity arises from the detection method and has proven to be very valuable for investigations by Klein and coworkers of films of water (where the friction can be very low) and simple liquids, as well as polymer melts and solutions.

Capacitance sensors have also been used in a two-dimensional friction measuring device with a reported sensitivity of 2 μ N [14]. The friction detecting capabilities were shown for confined hexadecane with a film thickness of two and three molecular layers.

Recent Technique Developments

Advances in Distance and Force Measurements

Many different improvements to the original techniques and setups have been proposed and implemented in the past decades. Significant steps have been made to automate the normal



Surface Forces Apparatus, Fig. 4 Schematic drawing of the surface force balance (SFB) designed by Klein and coworkers and configured for nanotribological measurements. The distance between the surfaces and the lateral

displacement of the upper surface is controlled with a sectored piezoelectric tube. The deflection of the shear force spring (S) is detected by an air-gap capacitor (Reprinted with permission from Ref. [13])

and lateral force measurement, and non-interferometric techniques have been introduced to measure the separation between the surfaces to enable the use of opaque substrates. Several designs for non-interferometric control and measurements of distance have incorporated piezoelectric bimorph elements [15, 16]. These devices have been used for thin-film viscosity measurements and studies of time- and rate-dependent adhesion in the laboratories of Israelachvili and Granick, respectively. Generally, piezoelectric elements are inadequate for long or quasi-static measurements because of their intrinsic drift and signal decay. The use of an ultrahigh impedance amplifier to lengthen the decay time of the bimorph sensor has to some extent helped overcome this difficulty. An entirely different method of distance detection is based on measuring the capacitance either between the silver layers on the back side of the mica sheets or between one plate of a capacitor attached to the

chamber of the instrument and one attached to the moving surface [17]. These techniques allow fast and accurate measurements of the displacement of the surfaces and eliminate the constraint of having to use transparent surfaces. However, when opaque surfaces are used, it is not possible to obtain a real-time, *in situ* image of the contact region during shear, which is one of the major strengths of the standard SFA technique.

Local Structural Information: Combination of SFA with Other Techniques

The data gathered in a conventional SFA experiment is the result of an average response of the confined film to shear and compression. Information on the molecular-level film structure during these processes would add significantly to the interpretation of the measured forces. Obtaining such information is very challenging because of the nature of the confined region: The response arises from a relatively small number of

molecules, which implies that any characteristic spectroscopic signal from this region will be of low intensity. The confined film is surrounded by thicker layers of materials that will give rise to spectroscopic signals of their own, typically much stronger than that from the film. Despite these experimental difficulties, interesting observations on film structure and molecular alignment have been made by combining SFA with several different techniques, as briefly described below. Further development of techniques suitable for studying the molecular properties of highly confined films can be expected to significantly improve our understanding of friction phenomena.

Extensions of standard MBI have been used to obtain structural information on the orientation and intermolecular interactions of optically active molecules in confined films [18]. Since the light adsorption of optically active (dye) molecules is enhanced by the multiple reflections in the optical cavity, very thin films can be studied, and local information in the contact area can be obtained.

Combining the SFA with other optical techniques has been limited by the presence of the reflective silver layers needed for the standard MBI approach. These layers strongly reduce the illumination of confined films, limiting the performance of spectrometric techniques. This has been successfully overcome by replacing the silver layers with multilayer dielectric coatings [19], with transparency to different wavelengths, so that information on the structure of ultrathin films under shear can be obtained. For example, the SFA has been combined with fluorescence correlation spectroscopy to measure the molecular diffusion coefficient in thin films within spots of submicron size, obtaining spatially resolved measurements [19]. A drawback is that fluorescent molecules have to be added to the film to be investigated. A reduction in the diffusion coefficient by about two orders of magnitude was found with confinement of films of simple liquids, and the diffusion coefficient was found to decrease when going from the edges toward the center of the contact region. Results have also been reported from combining the SFA with confocal Raman spectroscopy [19], which avoids the problem of the bulk contributing to the scattered

signal. Using multilayer reflective coatings, the geometry of the contact area could be monitored simultaneously with the Raman signal, showing the effects of shear on the orientations of molecules within the confined film.

The SFA has also been combined with x-ray diffraction measurements (the XSFA) [20]. Although this technique has thus far been limited to films thicker than 500 nm, the shear-induced alignment of liquid crystal molecules in confined films has been shown. Because of the size of the x-ray beam, the results represent an average over the contact area. Active research in this area is aiming at reducing the investigated region and film thickness, recently indicating that structural information on confined molecular films can be obtained with x-ray reflectivity [21].

The range of shear rates has been extended by incorporating a standard quartz crystal resonator as the lower surface in the SFA [22], giving a high oscillation frequency (MHz). A back-silvered mica piece was attached to the planar quartz crystal, and the top surface was mica glued to a half-spherical disk, so that a sphere-on-flat geometry could be obtained. Standard MBI was used to measure the diameter of the contact area and the radius of curvature of the sphere-on-flat contact geometry.

Mica and Beyond: Modified and Alternative Substrates

Muscovite mica is the most commonly used substrate in SFA experiments, because of its transparency and the relative ease by which large areas of atomically smooth, step-free sheets of uniform thickness (typically 2–5 μm) can be cleaved from larger crystals [5, 8]. The compressibility of mica is quite low, and it is chemically inert, i.e., in its native state, it does not expose functional groups to which chemical reactions can occur. On the one hand, this implies that its properties are quite well-known from one experiment to the other. On the other hand, a substantial effort is needed to modify mica for different purposes by physisorption or chemical reactions. The substrate plays a large role in the surface phenomena investigated by SFA, particularly in nanotribology, and unmodified (bare) mica is not necessarily

representative of many surfaces of engineering or biomedical interest. A large number of procedures have been developed for the purpose of modifying or replacing the mica surfaces with, for example, organic thin films, metals, or metal oxides to get surfaces with different chemical properties and surface energies (cf. Refs. [2, 4]).

Important considerations when modifying or replacing the mica substrates are the smoothness of the resulting surface and its transparency to light. An increase in roughness might change the standard single-asperity contact to a more complex, multi-asperity one. However, in many of the approaches described below, the roughness of the deposited layers can be controlled or modified at least to some extent, and in cases where rougher surfaces are formed, this allows investigations of the effects of surface roughness on the measured forces, which in itself is an important field of research. Electrochemical experiments in the SFA have shown that surface roughness affects the counterion distribution and thus the double-layer forces measured between charged surfaces [23] and have also demonstrated the possibility of electrochemical *in situ* modification.

Surface modifications by deposition of thin organic layers (often monolayers) can often be done so that the smoothness of the surface is close to that of the mica substrates themselves. Many different types of surface-active molecules (surfactants, lipids, block copolymers, proteins) have been deposited by self-assembly from solution or by Langmuir-Blodgett deposition [2, 4]. For example, in aqueous solution, the mica surface becomes negatively charged (because of solvation of K⁺ ions), and positively charged species spontaneously adsorb onto it. The friction response of the modified surfaces depends strongly on the properties of the adsorbed layers such as their surface energy, morphology, and inherent stiffness [2, 4]. The attachment of organic layers to the mica surface can be enhanced by chemical modification of the mica itself, for example, by water vapor plasma treatment. This enables the chemical binding of chlorosilane species to form robust, molecularly smooth, hydrophobic surfaces.

Mica surfaces have also been used as substrates for deposition of various inorganic materials, including metals and dielectrics. In many of these cases, the inertness of the native mica increases the risk of dewetting of the deposited layer, and some pretreatment of the mica (e.g., plasma treatment or the deposition of a separate adhesion layer) is necessary.

The analysis of the resulting FECO becomes more complicated because of the larger number of optical layers in the interferometer, but algorithms for this are available [11]. Thin deposited layers of materials such as silver, gold, platinum, silica, alumina, and zirconia with various film thicknesses and smoothness have been investigated [2, 4].

The interface between a supporting mica sheet and a material deposited on it has a roughness similar to that of the mica sheet itself. Because of the inertness of the mica, the bond between the mica and the deposited material is weak, and this can be utilized to form so-called template-stripped surfaces for use in the SFA [24]. The material of interest, for example, gold, is deposited on a freshly cleaved mica sheet to a desired thickness; its exposed surface is attached to another substrate (e.g., by gluing), and the mica sheet is peeled off with tweezers or with adhesive tape to expose a very smooth gold substrate that can be further modified through chemical reactions.

Substrates not based on mica or on a sacrificial mica substrate have also been developed and successfully investigated (cf. Ref. [4] and references therein): Among these, the earliest example is sapphire (aluminum oxide) single crystals grown from the vapor phase. More recently, silicon nitride surfaces have been formed by plasma-enhanced chemical vapor deposition onto rigid silica disks coated with a reflective layer [25]. A more easily prepared substrate, sheets of silica (quartz glass) or borosilicate glass [26], collected from a bubble with a thickness of a few to 10 µm formed by standard glassblowing techniques, has been used in a couple of studies. These flexible glass substrates, which are quite robust and have a roughness of only a fraction of a nanometer due to the surface tension of the glass when molten, can be silvered on their back side and used in a similar

manner to mica sheets. Surface modifications developed for glass and silica substrates (e.g., chemical bonding of silanes) can be readily applied to these substrates. Thin sheets of polymers have also been successfully prepared by casting and stretching for use in adhesion studies [27].

Cross-References

- [Atomic Force Microscopy](#)
- [Disjoining Pressure and Capillary Adhesion](#)
- [Friction Force Microscopy](#)
- [Nanotribology](#)

References

1. Israelachvili, J.N., McGuiggan, P.M.: Adhesion and short-range forces between surfaces. Part 1: new apparatus for surface force measurements. *J. Mater. Res.* **5**, 2223–2231 (1990)
2. Ruths, M., Israelachvili, J.N.: Surface forces and nanorheology in molecularly thin films. In: Bhushan, B. (ed.) *Springer Handbook of Nanotechnology*, 3rd edn, pp. 857–922. Springer, Berlin/Heidelberg (2010), and references therein
3. Luengo, G., Schmitt, F.-J., Hill, R., Israelachvili, J.: Thin film rheology and tribology of confined polymer melts: contrasts with bulk properties. *Macromolecules* **30**, 2482–2494 (1997), and references therein
4. Israelachvili, J.N.: *Intermolecular and Surface Forces*, 3rd edn. Academic, Amsterdam (2011), and references therein
5. Israelachvili, J.N., Adams, G.E.: Measurements of forces between two mica surfaces in aqueous electrolyte solutions in the range 0–100 nm. *J. Chem. Soc. Faraday Trans. I* **74**, 975–1001 (1978)
6. Parker, J.L., Christenson, H.K., Ninham, B.W.: Device for measuring the force and separation between two surfaces down to molecular separations. *Rev. Sci. Instrum.* **60**, 3135–3138 (1989)
7. Israelachvili, J., Min, Y., Akbulut, M., Alig, A., Carver, C., Greene, W., Kristiansen, K., Meyer, E., Pesika, N., Rosenberg, K., Zeng, H.: Recent advances in the surface forces apparatus (SFA). *Rep. Prog. Phys.* **73**, 036601 (2010), and references therein
8. Israelachvili, J.N.: Thin film studies using multiple-beam interferometry. *J. Colloid Interface Sci.* **44**, 259–272 (1973)
9. Stewart, A.M., Parker, J.L.: Force feedback surface force apparatus: principles of operation. *Rev. Sci. Instrum.* **63**, 5626–5633 (1992)
10. Tonck, A., Georges, J.M., Loubet, J.L.: Measurements of intermolecular forces and the rheology of dodecane between alumina surfaces. *J. Colloid Interface Sci.* **126**, 150–163 (1988)
11. Heuberger, M.: The extended surface forces apparatus. Part I. Fast spectral correlation interferometry. *Rev. Sci. Instrum.* **72**, 1700–1707 (2001), and references therein
12. Peachey, J., Van Alsten, J., Granick, S.: Design of an apparatus to measure the shear response of ultrathin liquids films. *Rev. Sci. Instrum.* **62**, 463–473 (1991), and references therein
13. Raviv, U., Tadmor, R., Klein, J.: Shear and frictional interactions between adsorbed polymer layers in a good solvent. *J. Phys. Chem. B* **105**, 8125–8134 (2001), and references therein
14. Qian, L., Luengo, G., Douillet, D., Charlot, M., Dollat, X., Perez, E.: New two-dimensional friction force apparatus design for measuring shear forces at the nanometer scale. *Rev. Sci. Instrum.* **72**, 4171–4177 (2001)
15. Israelachvili, J.N., Kott, S.J., Fetter, L.J.: Measurements of dynamic interactions in thin films of polymer melts: the transition from simple to complex behavior. *J. Polym. Sci. B* **27**, 489–502 (1989)
16. Dhinojwala, A., Granick, S.: New approaches to measure interfacial rheology of confined films. *J. Chem. Soc. Faraday Trans.* **92**, 619–623 (1996)
17. Stewart, A.M.: Capacitance dilatometry attachment for a surface-force apparatus. *Measure. Sci. Technol.* **11**, 298–304 (2000)
18. Mächtle, P., Müller, C., Helm, C.A.: A thin absorbing layer at the center of a Fabry-Perot interferometer. *J. Phys. II* **4**, 481–500 (1994)
19. Bae, S.C., Wong, J.S., Kim, M., Jiang, S., Hong, L., Granick, S.: Using light to study boundary lubrication: spectroscopic study of confined films. *Philos. Trans. A Math. Phys. Eng. Sci.* **366**, 1443–1454 (2008), and references therein
20. Idziak, S.H.J., Koltover, I., Israelachvili, J.N., Safinya, C.R.: Structure in a confined smectic liquid crystal with competing surface and sample elasticities. *Phys. Rev. Lett.* **76**, 1477–1480 (1996)
21. Seeck, O.H., Kim, H., Lee, D.R., Shu, D., Kaendler, I. D., Basu, J.K., Sinha, S.K.: Observation of thickness quantization in liquid films confined to molecular dimension. *Europhys. Lett.* **60**, 376–382 (2002)
22. Berg, S., Ruths, M., Johannsmann, D.: High-frequency measurements of interfacial friction using quartz crystal resonators integrated into a surface forces apparatus. *Phys. Rev. E* **65**, 026119 (2002)
23. Valtiner, M., Banquy, X., Kristiansen, K., Greene, G. W., Israelachvili, J.N.: The electrochemical surface forces apparatus: the effect of surface roughness, electrostatic surface potentials, and anodic oxide growth on interaction forces, and friction between dissimilar surfaces in solution. *Langmuir* **28**, 13080–13093 (2012)

24. Chai, L., Klein, J.: Interaction between molecularly smooth gold and mica surfaces across aqueous solutions. *Langmuir* **25**, 11533–11540 (2009), and references therein
25. Golan, Y., Alcantar, N.A., Kuhl, T.L., Israelachvili, J.: Generic substrate for the surface forces apparatus: deposition and characterization of silicon nitride surfaces. *Langmuir* **16**, 6955–6960 (2000)
26. Horn, R.G., Smith, D.T., Haller, W.: Surface forces and viscosity of water measured between silica sheets. *Chem. Phys. Lett.* **162**, 404–408 (1989)
27. Merrill, W.W., Pocius, A.V., Thakker, B.V., Tirrell, M.: Direct measurement of molecular level adhesion forces between biaxially oriented solid polymer films. *Langmuir* **7**, 1975–1980 (1991)

Surface Modeling of Ceramic Biomaterials

Marta Corno and Piero Ugliengo

Department of Chemistry, University of Torino,
Torino, Italy

Synonyms

2D-slab approach; Computational studies; Simulations of solid interfaces

Definition

A surface of a generic solid material can be considered as the topmost boundary layer of the material. If a crystalline solid is theoretically described as the infinite repetition of a unit cell in the three dimensions of space, the surface can be considered as the greatest defect of the crystal itself, since it interrupts the periodicity, *though a real crystal is a macroscopic object made of finite atoms*. Indeed, physical and chemical phenomena are in a large number of cases occurring at the interface between a solid material and the environment, i.e., at its surface. To further highlight the relevance of surfaces in science, it is sufficient to mention that the 2007 Nobel Prize in Chemistry was awarded to Gerhard Ertl “for his studies of chemical processes on solid surfaces”[1]. To this respect in the last decades, computational methods have become relevant in the study of surface properties. Therefore, surface modeling has acquired a great interest, both theoretically and for practical applications in the chemical industry, e.g., in catalysis, adsorption, corrosion, and oxidation, and in the biomedical field. Especially in this last research area, complex mechanisms are involved in the interaction among material surfaces and body biological fluids. Surface modeling can then help in the progress of scientific and technological knowledge for the design of improved biomaterials to be integrated within the human body.

Surface Free Energy and Chemical Potential at Nanoscale

- Surface Energy and Chemical Potential at Nanoscale

Surface Hopping

- Theory of Nonadiabatic Electron Dynamics in Nanomaterials

Surface Loss in Micromechanical/Nanomechanical Resonators

- Surface Dissipations in NEMS/MEMS

Surface Loss in NEMS/MEMS

- Surface Dissipations in NEMS/MEMS

Surface Micro-patterning

- Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (P μ CP) Technique

Introduction to Surface Modeling

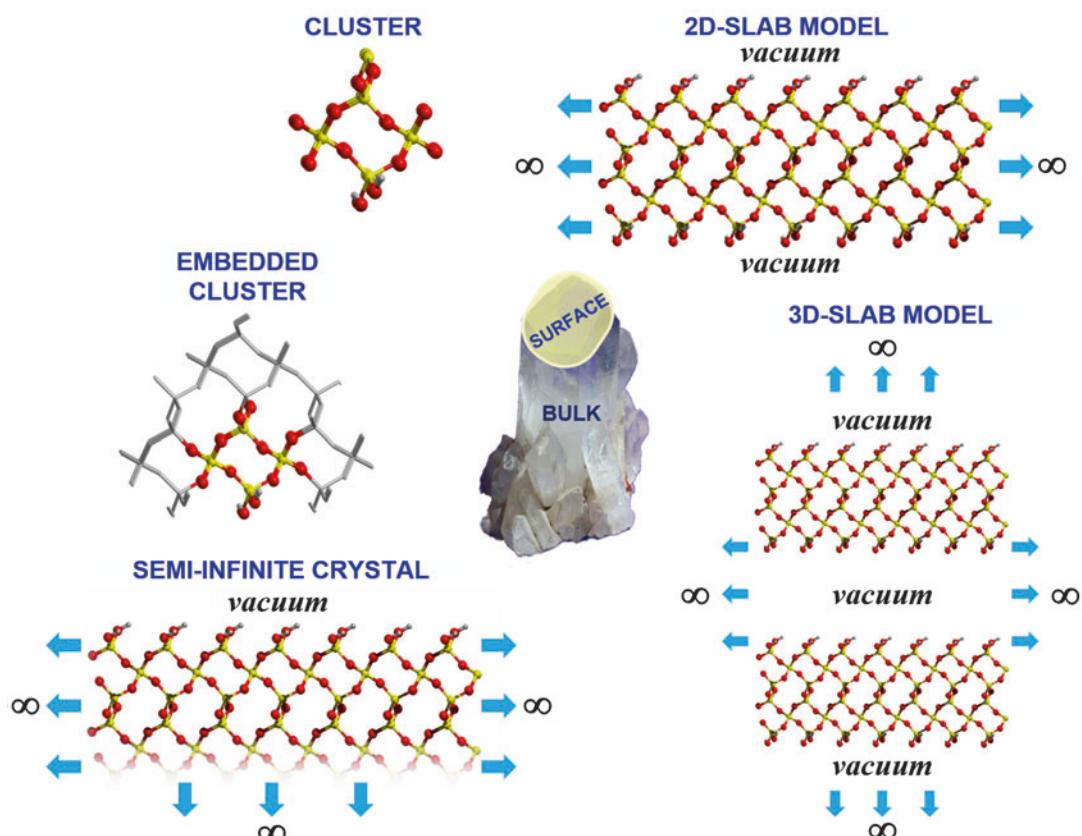
Quantum mechanics is the fundamental physic framework for studying processes occurring at surfaces of solid materials, including both static and dynamic phenomena. For the latter, the treatment with a classical or semiclassical mechanics approach can be also mentioned, characterized by the development over the years of ad hoc force-fields. By means of the increasingly accurate computational techniques, experimental findings have become reproducible by simulations and often even clearer to be interpreted due to the synergic use of theory and experiments. In this sense, it is worth mentioning that Density Functional Theory (DFT) methods have been and currently are successfully coupled with experimental surface

science techniques, as scanning tunneling microscopy (STM), temperature-programmed desorption (TPD), and X-ray diffraction, among others, to gain a deeper knowledge of surface properties for materials of different chemical nature (metals, oxides, nanoparticles, etc.) [2].

The complete and rigorous treatment of quantum-mechanical methods for computing surface properties is outside the scope of this essay. A more general overview of the most frequently used approaches to surface modeling will be exposed in the following sections.

Main Approaches to Surface Modeling

For modeling surfaces three main approaches are possible: cluster (or embedded cluster), semi-infinite crystal, and slab models (Fig. 1).



Surface Modeling of Ceramic Biomaterials,
Fig. 1 Main approaches to surface modeling for the case of quartz. *Blue arrows* and the infinite symbol indicate periodicity; atoms are colored as follows: silicon in *yellow*,

oxygen in *red* and hydrogen in *light gray*. In the embedded cluster model, atoms represented as gray sticks are those treated with a lower degree of accuracy with respect to the full colored ones

In the *molecular cluster* approach, a piece (cluster) of atoms belonging to the extended surface and representative of the specific site to be simulated is cut out from the surface itself. Obviously, the larger the size of the cluster the better the similarity with the original surface. Within this model no periodicity is considered, so the translational symmetry is lost and the point group to which the cluster belongs is the new reference for symmetry operations. Using this approach, caution has to be paid due to the introduction of spurious effects originated by the limited size and the need to add extra atoms to cap the unfilled valences at the cluster boundaries. Usually, hydrogen atoms are used to cover the cluster frontier. Because of the intrinsic limitations of the cluster model for representing surfaces, a number of factors must be taken into account to evaluate the reliability of the obtained results. First of all, the size has to be carefully chosen, by analyzing the dependence of the cluster physicochemical properties on the cluster size. Moreover, other fundamental requisites are electroneutrality and the same stoichiometry, average atomic coordination, and symmetry elements as in the bulk. To overcome the cluster size limitations, an alternative technique consists in embedding the cluster in an external field which mimics the missing bulk crystal environment. In this way the whole system is divided into subsystems, which are treated with different levels of theory and consequently accuracy (e.g., by adopting Green function techniques). The embedding technique is also commonly adopted when handling the simulation of defects in solids.

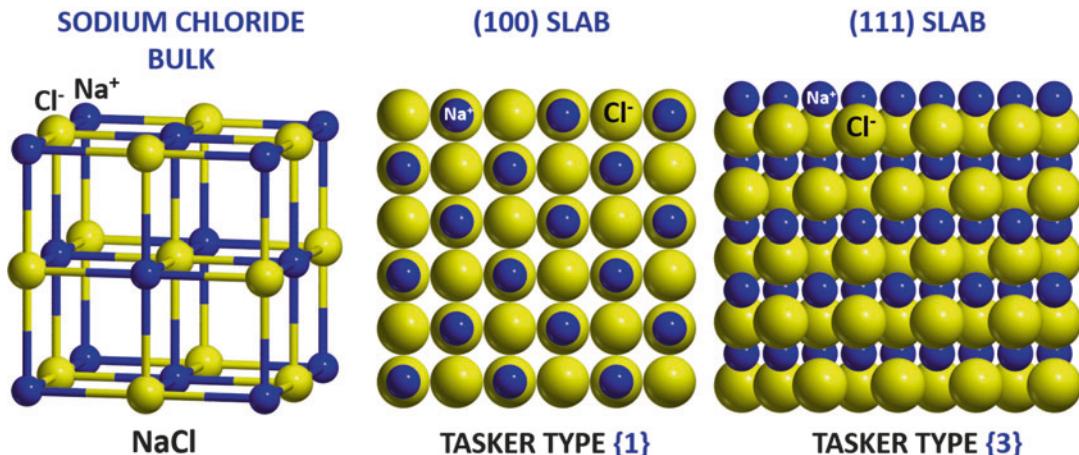
In the second approach to surface modeling, the *semi-infinite crystal* approach, two semi-infinite crystals are obtained by cutting the 3D crystal through a selected crystalline plane. The periodicity is maintained in 2D by the semicrystals, while is lost along the perpendicular z direction and the problem is reduced to an infinite + finite problem, that is the few layers close to the surface constitute a finite subsystem connected to the semicrystal with known electronic structure. This approach requires the theoretical framework of the Green functions and in the past has been successfully applied for alkali metals.

In the third and most common approach, the *slab model*, the bulk structure is cut to extract a layer of finite thickness exposing two surfaces. However, the first operation is the choice of the reference surface plane for cutting the crystal bulk. Conventionally, each surface is indicated by three integers, known as Miller indices (hkl), which identify the family of crystallographic planes to which the surface belongs. The most probable and stable surfaces are usually those characterized by the lowest Miller indices, as they envisage surfaces with the highest atomic density. Usually, the kind and extension of surfaces are derived from the experimental crystalline habit, or from the analysis of experimental images recorded by means of the transmission and scanning electron microscope.

The slab model can be implemented following two different schemes: (i) a 2D slab envisaging true vacuum space at top/bottom of the slab and extending through infinity, while keeping two-dimensional periodicity within the slab itself or (ii) a series of slabs is repeated along the direction perpendicular to the slab plane and separated by vacuum gaps, spaced apart enough to avoid spurious interactions among replicas. This latter case is indeed a pseudocrystal in which the unit cell to be repeated in 3D contains the slab of a given thickness plus the vacuum gap. The choice between the two schemes is associated to the choice of the adopted basis set functions used to represent the wave functions of the Schrödinger equation to be solved within the approximate quantum-mechanical method [3]. For instance, when plane waves functions are adopted (functional form $\phi_k(r) = \exp^{ik \cdot r}$), only the 3D slab model is suitable, due to the inherent periodicity of plane waves. Conversely, computer programs based on localized Gaussian-type basis sets can adopt both models [4].

Surface Stability

When considering ionic crystalline materials, the stability of the corresponding modeled surfaces follows a classification proposed by Tasker, stating that three categories exist: (i) type 1 slab, with neutral planes or layers containing both anions and cations (same stoichiometry of the parent



Surface Modeling of Ceramic Biomaterials,

Fig. 2 Example of Tasker type surface models for sodium chloride (NaCl): leftmost side, the structure of the corresponding sodium chloride bulk unit cell; center, lateral view of the NaCl (100) slab, characterized by six

neutral layers, with both anions and cations in the two outermost layers (type 1); and right side, lateral view of the NaCl (111) slab, exhibiting ten alternating charged layers of anions and cations, with the two outermost conferring a net dipolar moment across the slab (type 3)

crystal); (ii) type 2 slab with charged planes or layers arranged symmetrically so that no net dipole moment is generated across the surface; and (iii) type 3 slab with charged planes or layers imparting a net dipole moment across the surface [5]. The definition of type 3 dipolar surface brings about an inherent instability due to the buildup of an infinite dipole moment with the increasing slab thickness ultimately breaking the system apart. Figure 2 reports an example of surfaces for sodium chloride, classified according to Taker types. One should be aware that the Taker's classification assumes a truly ionic nature of the material in which no charge transfer or polarization is permitted. Obviously, this is not the case when the system is treated quantum mechanically and the nature of the material is not entirely ionic. In the following, the case of the dipolar (001) hydroxyapatite surface will be described, in which the role of polarization and charge transfer allow the existence of this surface up to thickness of hundredths of nanometers.

For both slab and multislab models, one crucial parameter to carefully evaluate is the number of layers forming the slab, i.e., its thickness. Since the slab is cut out from an infinite crystal, the thicker is the surface (or larger is the number of

layers), the closer the slab will be to the “real” material. The energy needed to create the surface from the bulk (surface energy) is also dependent on the slab thickness. In principle, the correct thickness for a given slab implies a converged value of the surface energy with the slab thickness. In practice, one should compromise between accuracy of the converged surface energy (improving with the slab thickness) and cost of the calculation (increasing with the slab thickness).

The surface energy (E_s , γ or σ depending on the chosen formalism) is the energy required to cut the slab out of the bulk. Lower is the surface energy, more stable is the considered surface. The general definition of the surface energy is the following

$$E_{\text{surf}} = \frac{(E_{\text{slab}}^n - n \cdot E_{\text{bulk}})}{2A} \quad (1)$$

Where E_{slab}^n is the energy of an n -layered slab, E_{bulk} is the energy of a formula unit of the material in its bulk, n is the number of formula units in the slab model, A is the surface area, doubled due to the presence of two limiting surfaces. This definition is for symmetric slabs only, where the two limiting surfaces are identical. The same

definition is also adopted for nonsymmetrical cases, in which the E_{surf} is usually considered as an average surface energy value. The surface energy is a positive quantity, otherwise the bulk structure would exfoliate, and is expressed in units of J/m^2 . By increasing the slab thickness ($n \rightarrow \text{infinite}$), the surface energy should converge to a reference value and this check, as already mentioned, is essential for comparing the relative stability of different crystal surfaces, as the E_{surf} convergence depends strongly on the (hkl) plane. The definition of E_{surf} by Eq. 1 involves unit cell energies for systems of different dimensionality (2D vs 3D). Depending on the computational code, care should be taken to ensure that these energies are computed with the same accuracy for 2D and 3D cases. To mitigate the problem, the E_{surf} can be computed by alternatives formulas:

$$E_{\text{surf}} = \frac{\left[E_{\text{slab}}^n - \frac{n}{2} (E_{\text{slab}}^n - E_{\text{slab}}^{n-1}) \right]}{2A} \quad (2)$$

$$E_{\text{surf}} = \frac{(E_{\text{slab}}^n - E_{\text{bulk}}^n)}{2A} \quad (3)$$

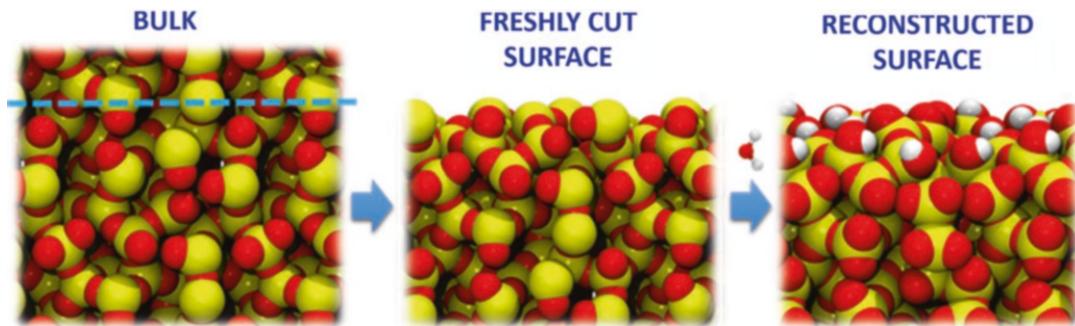
In both equations, the idea is to compare unit cell total energy values computed in the closest computational conditions. For Eq. 2, only energies from slab models are considered, while in Eq. 3 the unit cell of the bulk is built as a super cell of the reference cell to ensure that the crystal vector along the cut has the same module of the slab thickness.

Equations 1, 2, and 3 are only valid for chemical composition of the surface slabs which are an integer multiple n of the bulk unit cell content (stoichiometric surfaces). When dealing with nonstoichiometric surfaces formulas Equations 1, 2, and 3 are no longer applicable. The E_{surf} can be arrived at by a more complicated approach which takes into account the chemical potential of single ions or species needed to keep the system in thermodynamic equilibrium. In this way, E_{surf} will depend on the reference state with respect to which chemical potential values have been chosen.

Surface Relaxation and Reconstruction

According to the above definition, it is possible to list a ranking of stability for various surfaces of the same material, characterized by diverse Miller indices, and compare it to experimental data, if present. The ranking of surface relative stability can give information about the crystal morphology on a purely thermodynamic ground. Kinetic effects frequently dominate the final morphology of a real crystal and should be taken into account by different methods which will not be addressed here.

Nevertheless, the (hkl) indices only define the direction perpendicular to which the slab is cut; however, the specific position of the cut plane is somehow dictated by minimizing the number of chemical bonds to be cut at the surface. This process is not straightforward, because as a function of the chemical nature of the material (ionic, metallic, molecular, or covalent) different strategies are needed to cope with the unfilled valencies resulted by the bonds cut. For ionic and metallic surfaces, there is no need to add extra atoms provided that, for ionic crystals, the system is kept electroneutral. For molecular crystals, the cut maintains the molecular integrity of the crystal constituents resulting in a rather rough surface. For these cases, the geometry of the final slab should be fully relaxed by minimizing the forces acting on all atoms. This process is called “relaxation” and is indeed very crucial to compute a correct E_{surf} . Its importance is obvious: the most exposed atoms/ions or molecules will move inwards the slab to maximize the number of interactions with the bulk like atoms. For instance, surface cations in ionic crystals will move inwards to increase the favorable Coulombic attraction with the underneath anions. For molecular crystals, the outmost molecules will relax to establish the missed interactions due to the cut through new H-bonds and dispersive interactions with the nearby molecules. The most difficult case is for covalently bound crystals. In that case, there is no way to avoid unfilled valences at the newly created surface. For a homolitic cut simulating a process occurring in ultrahigh vacuum, the unpaired electrons will pair to reconstruct new chemical bonds. This is the case of the silicon



Surface Modeling of Ceramic Biomaterials, Fig. 3 Example of surface restructuring by water: the case of amorphous silica (silicon yellow, oxygen red, hydrogen white)

(100) surface. This kind of process causes large relaxation energy in the reconstructed surfaces. When covalent crystals are formed in presence of water moisture, or directly from water solution, the surface is chemically reconstructed by the reaction of water molecule and the unsaturated atoms at the surface, resulting in the formation of hydroxyl OH groups. A concrete example of modeling surface reconstruction by water is the case of silica, as for instance in the amorphous phase, as illustrated by Fig. 3. For this structure, defining a surface means cutting the Si-O bonds homolitically, leaving Si and O at the surface which will be converted to SiOH and OH groups by reacting with water molecules. Indeed, the $-SiOH$ groups, silanols, have been detected experimentally at silica surfaces by a variety of spectroscopic techniques [6]. This is an example of surface reconstruction, in which, at variance with the silicon case, the chemical composition is changed with respect to the bulk. Other more extreme examples of surface reconstruction may imply the exchange of ions from surface top to bottom and vice versa, with the aim of stabilizing polar structures of Tasker type 3.

Both relaxation and reconstructions are essential steps in surface characterization. Relaxation can be dramatic if too thin slab models are adopted to simulate a given surface. Reconstruction always implies the need of a careful relaxation due to the change in chemical environment of the newly created functionalities.

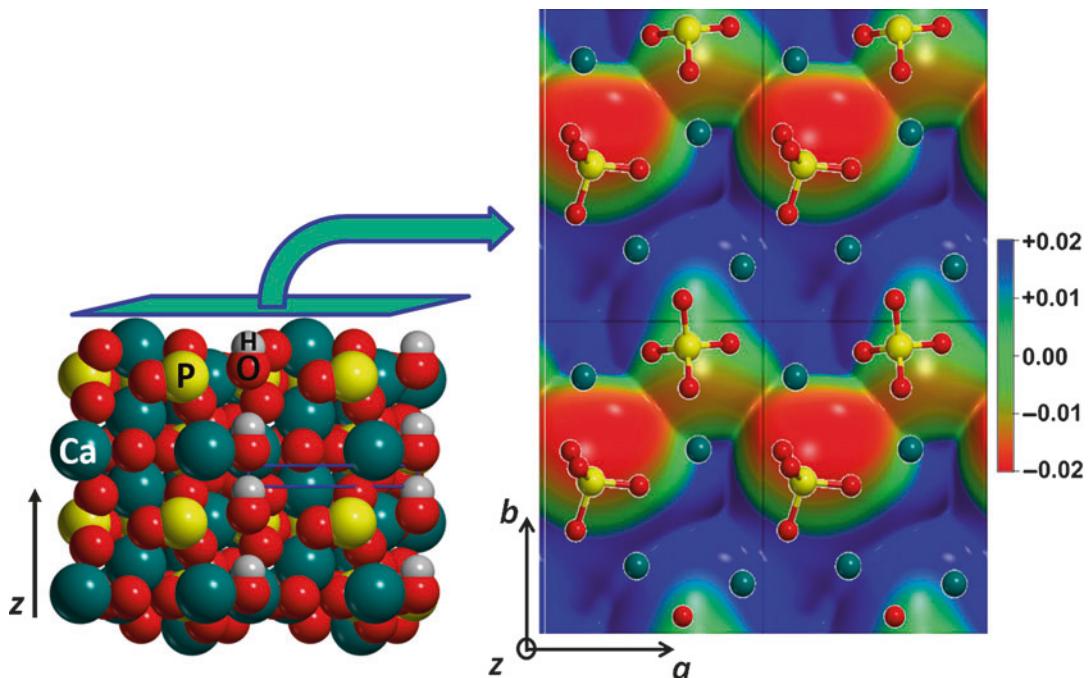
Another degree of complexity in surface modeling is the definition of defective surfaces,

since real surfaces are neither fully stoichiometric nor are perfectly flat. The first class of defects envisages substitution of chemically related elements (for instance Na by K) or atom vacancies. The second class of defects includes the presence of steps and kinks. Typically, vicinal surfaces with high Miller indices, which are those surfaces cut at a relatively small angle to one of the low index surfaces, are used to define stepped surfaces. The chemical reactivity of the resulting surface models is often studied towards molecules adsorption.

Simulation of Surface Adsorption Processes

Among physicochemical properties of surfaces which can be derived by quantum-mechanics calculations, one of the most informative and useful for further adsorption studies is the electrostatic potential at the surface. Furthermore, by mapping the electrostatic potential on the electron charge density, a clear picture of the distribution and presence of negative, positive, and neutral potential values is obtained. Indeed, as displayed in Fig. 4 for the (001) hydroxyapatite surface (*vide infra*), colored zones are visible on the maps, and the color codes identify negative potential ones as red, positive as blue, and neutral as green. By performing this kind of analysis, an immediate picture of the different electrostatic fields experienced by a molecule approaching the surface is gained.

The simulation of electrostatic properties at surfaces becomes of great importance for the design of proper models of adsorption processes happening on specific faces of the examined



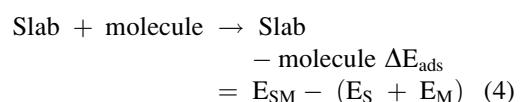
Surface Modeling of Ceramic Biomaterials,

Fig. 4 Electrostatic potential mapped on the electronic density for the (001) hydroxyapatite surface. *Left side:* side view of the hydroxyapatite slab (atoms as van der Waals spheres, colors: calcium cyan, phosphorous yellow,

oxygen red, hydrogen light gray; blue lines identify the unit cell borders); *right side:* top view of the electrostatic potential mapped on the electron density isosurface (isovalue: $10^{-5}e$, color scale shown in the picture)

material, considering both physisorption and chemisorption. Indeed, the electrostatic complementarity principle can help to define a reasonable starting adsorption geometry. For example, acidic moieties (COOH , OH , etc.) of adsorbed molecules orient towards electrophilic zones of the surface and vice versa for basic moieties (CO , NH_2 , etc.) in adsorbed molecules. The simulation of these processes can be conducted via the cluster model (the adsorbed molecule is part of the cluster system) or the slab model (in both 2D and multislab, though the first approach is more natural).

Usually, in ab initio static calculations, the starting geometry of an adsorbed molecule/surface structure is optimized with the chosen computational code and the optimized total energy is compared to the corresponding values for optimized isolated surface and molecule. So doing, the interaction energy per unit cell associated to the adsorption reaction of one molecule onto a surface can be computed as follows



for a reference reaction in the gas phase where SM is the slab/molecule adduct, S and M the free slab and molecule, respectively. All three energy values are referred to optimized structures and are negative, implying a negative value for ΔE_{ads} for favorable adsorption processes.

Instead of the interaction energy, in studying adsorption, the binding energy (BE) can be considered, defined so that $\text{BE} = -\Delta E_{\text{ads}}$. The more negative the interaction energy ΔE_{ads} or the most positive the binding energy BE is, the stronger the interaction between the molecule and the surface. If the molecule interacts with its replica in the nearby unit cell, also the lateral interactions contribute to the adsorption energy and must be included in the ΔE_{ads} estimation.

In the 2D slab model approach with localized Gaussian functions, the basis set is never complete, so that the Basis Set Superposition Error (BSSE) biases the interaction energy. The BSSE derives by summing and subtracting energy values computed with different set of functions in the three cases of bare surface, isolated molecule and adsorbent-adsorbate complex. If not taken into account, the BSSE can heavily affect the ΔE_{ads} (by giving too negative values), leading to the misinterpretation of computed results for the adsorption phenomenon. A convenient way to correct the binding energy is the counterpoise method (CP), in which energy calculations are repeated for the slab and molecule by including the extra basis set functions of each other.

Simulation of Bioceramic Surfaces

In a very general definition, biomaterials are those natural or synthetic materials meant to be in contact and interact with biological systems, and eventually able to substitute, repair, or reconstruct damaged body tissues [7, 8]. Among the numerous classes of biomaterials, bioceramics are considered biocompatible and osteoconductive. Physical and chemical interactions between complex biological systems and inorganic bioceramics must be investigated to improve the knowledge for designing new improved prosthetic materials.

Surface modeling is a useful tool to predict and interpret the experimental measurements, especially in describing, at an atomistic level, the interaction between the ceramic biomaterials and biological molecules.

In the following section, an example of surface modeling applied to a common bioceramic-hydroxyapatite, will be reported. All the theoretical concepts and quantities discussed in the previous part will be used here in practical examples.

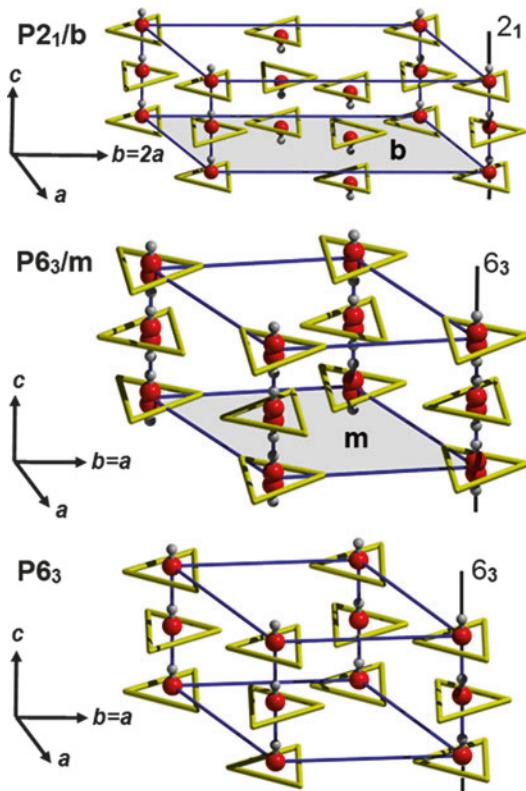
How to Model Hydroxyapatite Surfaces

Hydroxyapatite [HA, $\text{Ca}_{10}(\text{PO}_4)_6(\text{OH})_2$] is the main constituent of the inorganic phase in bone and tooth tissues and has been widely studied as a

biomaterial, mainly in combination with other materials for compensating its brittleness.

HA is also involved in the bioactivity mechanisms of other ceramic biomaterials, since according to the so-called Hench mechanism the crucial step for the integration of the inorganic prosthetic material within the human living tissues goes through the formation of a layer of carbonated hydroxyapatite. The role of this biomaterial's surface in the complex chemical and biological reactions of molecular and cellular recognition is of paramount importance. For that reason, a large number of experimental and theoretical studies have been performed, and the issue of surface modeling has become essential for the simulation and the understanding of the interaction processes, also when hydroxyapatite nanoparticles are concerned.

As described in the previous theoretical section, a surface is modeled starting from the deep knowledge of the corresponding bulk structure, irrespective of the chosen modeling technique (cluster, semi-infinite crystal, slab, or multislab model). In case of the hydroxyapatite crystal, the bulk structure, as revealed by X-ray and neutron scattering analysis, can be found in nature in two polymorphs, hexagonal and monoclinic. The distinctive feature of HA is the presence of columns or channels of hydroxyl groups (OH), aligned parallel to the crystallographic *c* axis. Indeed, these OH columns are inserted in the pattern where calcium and phosphate ions are disposed in a hexagonal fashion for both crystal forms. In the last 20 years, it has been established that the HA structure suffers of a large scale proton disorder, meaning that the orientation of the OH groups can vary, with the hydrogen atom point upwards or downwards with respect to the reference *c* axis. The difference among the monoclinic and the hexagonal phases is precisely how the OH groups are oriented with respect to the *c* axis orientation. In the monoclinic case, neighbor OH columns exhibit antiparallel orientations, while in the hexagonal one the OH orientation changes within the same column. In order to computationally treat the hexagonal form, which is the one mostly involved in biomineralization processes, the conventional way is to select a specific orientation of the OH



Surface Modeling of Ceramic Biomaterials,
Fig. 5 Different HA structures: monoclinic $P2_1/b$ (top), experimental hexagonal $P6_3/m$ (middle), and theoretical adopted hexagonal $P6_3$ (bottom) unit cells. P, O, and Ca atoms not relevant for the scheme were omitted for clarity of representation. Ca ions are at the vertices of triangles around each hydroxyl group. Reproduced from Ref. 9 with permission of the PCCP Owner Societies

groups, so fixing the hydrogen positions inside the crystal unit cell, as shown in Fig. 5. In this way, the HA crystal structure is a feasible model for theoretically investigating its physicochemical properties, such as geometrical, electronic, mechanical, and vibrational properties, also with the intent to assess the model by comparing calculated values to experimental data.

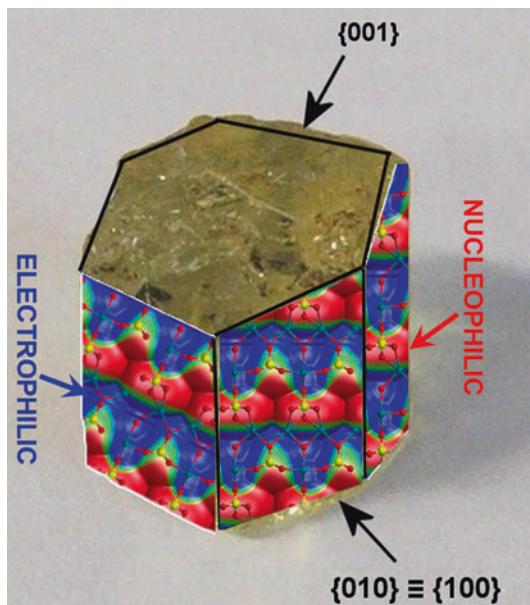
(001) and (010) HA Surfaces

Once the hydroxyapatite crystal bulk model has been validated, the first step towards surface designing has required the inspection of the crystalline habit, which is informative of the type and extension of specific faces. In the HA hexagonal

case, crystallites have needle-like morphology, and the most exposed face is the stoichiometric (010), which is equal to the (100) in the hexagonal system. Furthermore, it is known from the literature that the biomineralization process of bones and tooth enamel involves the (001) surface of hydroxyapatite, which is less exposed but directly interacting with the collagen fibers in bone and enamel growth [9]. In the framework of DFT calculations and taking as a reference the slab model techniques, several computational studies have concerned these two HA surfaces, not only per se but also in interaction with (bio)molecules [9–12].

The (001) plane represents the most natural cut of the bulk structure and does not require any covalent bond breaking. Nevertheless, the aligned OH columns build up a finite dipole moment across the (001) slab. As discussed, this could compromise the surface stability with the increasing thickness, i.e., the surface energy would increase instead of reaching a constant value at convergence with the slab thickness, causing the surface to collapse. The issue is quite complicated, since this slab has a chemical nature intrinsically different from the Tasker type 3 surfaces, where anions and cations were alternating in charged layers (*vide supra*) [5]. The careful inspection of the variation with thickness of a large number of properties, such as dipole moment along the cut direction, band gap, electrostatic potential at surface, surface formation energy, and others, proved only a moderate increase of the dipolar character, without signs of electronic instability. Recent calculations have revealed that for a slab thickness of about 40 nm, that is the real size of nanocrystals found in collagen, the surface is still stable (dipole moment lower than 1 Debye), without any sign of metallization typical of the classical type 3 surfaces, e.g., electronic band gap closure. This (001) HA surface indeed can be seen as an example of ferroelectric surface, whose stability is granted by a delicate balance between the field due to the OH groups and the counter-polarization due to calcium and phosphate ions, the latter free to rotate to minimize the dipole moment across the slab [13].

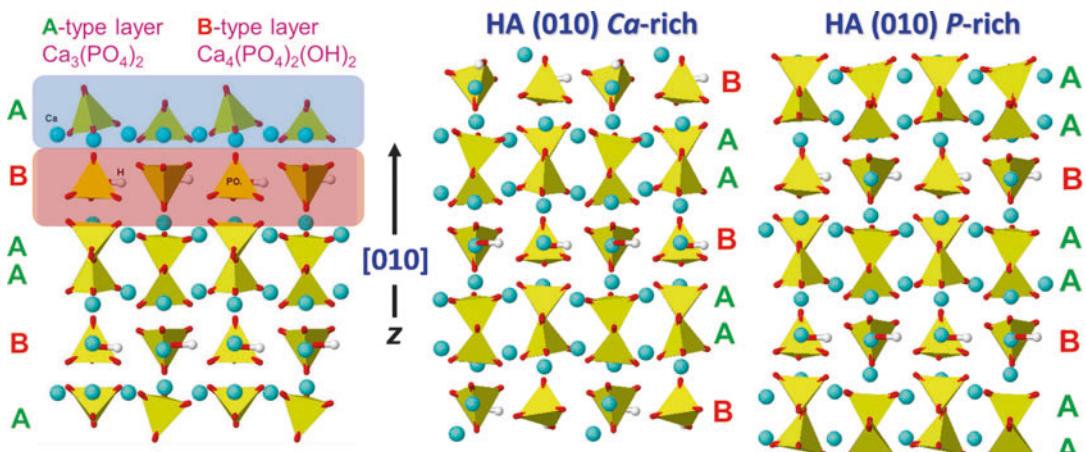
As for the most extended surface, the (010) one, the hydroxyl groups are aligned perpendicularly to the direction of the cut, and the surface is free from the problems affecting the (001) one.



Surface Modeling of Ceramic Biomaterials,
Fig. 6 Electrostatic potential mapped on the electronic density for the (010) hydroxyapatite faces, superimposed on a picture of the real HA crystal

Comparing the topology of the two surfaces, different exposed ions are available at surface for interacting with molecules, as it can be seen in Fig. 6 from the electrostatic potential mapped on electron density superimposed to a picture of the HA crystal for sake of clarity. Red zones correspond to negative values of the electrostatic potential and are due to the phosphate anions, while the blue zones are associated to calcium ions, i.e., positive values. By applying an electrostatic complementarity principle, adsorption processes at the two surfaces have been simulated for several molecules to probe specific surface physicochemical features and to improve the knowledge of reaction mechanisms within the body.

For hydroxyapatite, as for other materials, also nonstoichiometric surfaces are relevant, as experimentally shown for the (010) case [14]. Indeed, considering the [010] direction of the cut, two layers can be distinguished based on chemical composition: $\text{Ca}_3(\text{PO}_4)_2$, called A-layer and $\text{Ca}_4(\text{PO}_4)_2(\text{OH})_2$, called B-layer. The stoichiometric bulk ratio is A_2B , and in Fig. 7 these layers are displayed, with the indication of the three different terminations for the (010) surface family: (1) stoichiometric surface ($\text{Ca}/\text{P} = 1.67$); (2) nonstoichiometric with a ratio Ca to P larger than the stoichiometric one, called Ca-rich surface ($\text{Ca}/\text{P} = 1.71$); and (3) nonstoichiometric with a

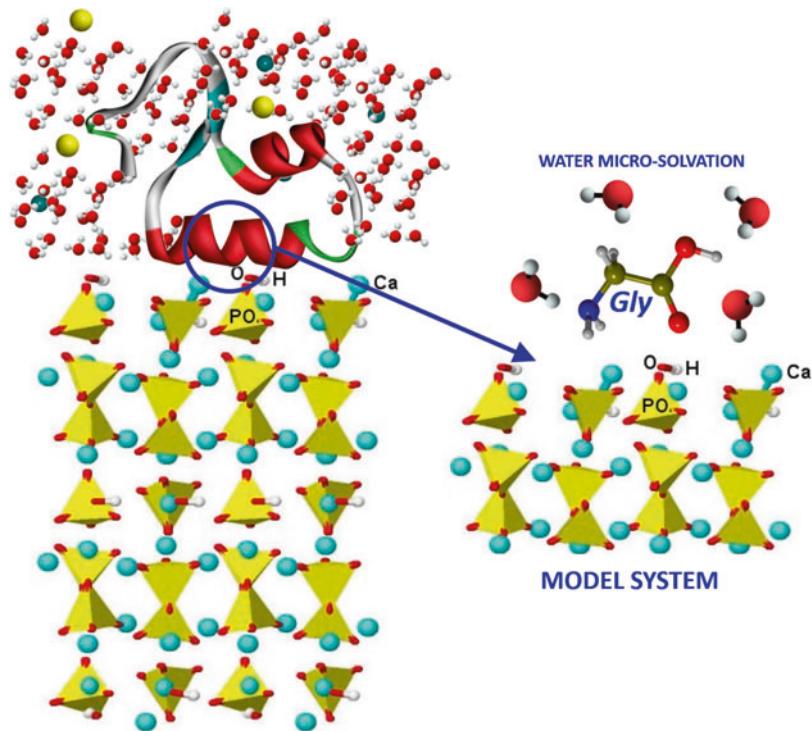


Surface Modeling of Ceramic Biomaterials,
Fig. 7 Graphical representation of: (left) the A and B layers alternating along the [010] direction in the HA

bulk structure (middle) the HA (010) nonstoichiometric Ca-rich surface and (right) the HA (010) nonstoichiometric P-rich surface

Surface Modeling of Ceramic Biomaterials,

Fig. 8 Simplified approach to the study of solvated protein adsorption on hydroxyapatite surfaces (*left*): only a few water molecules microsolvate the simplest amino acid, glycine, on the top of a hydroxyapatite slab of finite thickness (*right*)



Ca/P ratio lower than the stoichiometric, called P-rich surface ($\text{Ca}/\text{P} = 1.62$). As abovementioned, the issue of computing surface energy for nonstoichiometric surfaces is not trivial. The thermodynamic [15, 16] treatment has been applied, by considering as a reservoir the chemical potential value of pure calcium phosphate and calcium hydroxide. Thus, by combining and varying the chemical potential values of the two reference solid phases, a comparison can be obtained among the two nonstoichiometric surfaces in term of surface energy, and, ultimately, their relative stability.

HA Surfaces in Interaction with Biomolecules

In this section, only a brief mention to two examples of adsorption processes on the above described HA surfaces will be reported.

The first example aims at unveiling if water molecules compete with biomolecules (i.e., proteins) for the interaction with hydroxyapatite surfaces. In principle, the study of water solvation should be addressed by means of ab initio molecular dynamics techniques, considering a large number of molecules and requiring extremely heavy computational

resources. However, to reduce the complexity of the system and obtain accurate answers, a microsolvation approximation can be followed, that is modeling a few water molecules in interaction with specific functionalities of the selected biomolecule. In detail, the adsorption of the simplest amino acid, glycine (Gly), was simulated on (001), (010) – stoichiometric and nonstoichiometric – hydroxyapatite surfaces at first in the gas phase, to establish the most favorable interactions (see Fig. 8). For the case of (001) surface, water molecules were added on the most stable HA and Gly-HA structures and the results confirmed experimental findings that claimed a direct contact of the amino acid with the hydroxyapatite surface. Calculations were able to simulate the displacement of water by the incoming glycine molecules [17]. Obviously, the mentioned approach is characterized by drastic simplifications, so results must be taken with caution, but constitute a preliminary step towards a more complex design of the system, for instance with a larger number of molecules and a small peptide instead of a single amino acid.

The second example goes in this direction, since the interaction of a small oligopeptide, a polyglycine

of 12 Gly, was simulated on the surface of a monoclinic HA, in the gas phase. The goal of the computational study was to investigate if the interaction of a protein with HA surfaces could change the protein conformation, so influencing the biochemical processes occurring when an inorganic prosthetic material is inserted in the body. Interestingly, the oligopeptide had to be modified by substituting Gly residues in amino acids with a higher electrostatic affinity to the HA surface (lysine and glutamic acid). Without describing all the details, it is worth mentioning, as the most relevant result of this study, the stabilizing effect of HA surface for the α -helix conformation of the mutated oligopeptide, with respect to the folded conformation. In this sense, calculations would compare to a classical example of the statherin protein, which changes its role when in different conformations. Indeed, in the folded state, it prevents calcium phosphate to precipitate, while due to the contact with the HA surface of the teeth enamel, a partial unfolding has been observed [18].

Conclusion

Surface modeling has gained in the last decade great interest due to the increasing computational power, especially in the biomedical field. Understanding the atomistic details of reaction mechanisms at surface of inorganic biomaterials can be fundamental for an improved knowledge and design of the new materials. By means of slab modeling, the physicochemical properties of real 2D layers are computed and can be exploited for subsequent interaction with selected molecules. As a specific example, the simulation of hydroxyapatite surfaces represents a starting point for more complex studies of interaction with proteins towards the simulation of biomineralization processes.

Cross-References

- [Ab Initio DFT Simulations of Nanostructures](#)
- [Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling](#)

- [Computer Modeling and Simulation of Materials](#)
- [Molecular Dynamics Simulations of Nanobiomaterials](#)

References

1. Ertl, G.: Reactions at surfaces: from atoms to complexity (Nobel lecture). *Angew Chem Int Ed* **47**, 3524–3535 (2008)
2. Sholl, D.S., Steckel, J.A.: *Density Functional Theory*. Wiley, Hoboken (2009)
3. Jensen, F.: *Introduction to Computational Chemistry*, 2nd edn. Wiley, Chichester, England (2006)
4. Dovesi, R., Civalleri, B., Orlando, R., Roetti, C., Saunders, V.R.: Ab initio quantum simulation in solid state chemistry. *Rev Comp Chem* **21**, 1–125 (2005)
5. Tasker, P.W.: The stability of ionic crystal surfaces. *J Phys C Solid State Phys* **12**, 4977–4984 (1979)
6. Rimola, A., Costa, D., Sodupe, M., Lambert, J.-F., Ugliengo, P.: Silica surface features and their role in the adsorption of biomolecules computational modeling and experiments. *Chem Rev* **113**, 4216–4313 (2013)
7. Williams, D.F.: *Williams Dictionary of Biomaterials*. Liverpool University Press, Liverpool (1999)
8. Ratner, B.D., Hoffman, A.F., Schoen, F.J., Lemons, J.E.: *Biomaterials science*. Elsevier, New York (2013)
9. Corno, M., Rimola, A., Bolis, V., Ugliengo, P.: Hydroxyapatite as a key biomaterial: quantum-mechanical simulation of its surfaces in interaction with biomolecules. *Phys Chem Chem Phys* **12**, 6309–6329 (2010)
10. Almora-Barrios, N., Austen, K.F., de Leeuw, N.H.: Density functional theory study of the binding of glycine, proline, and hydroxyproline to the hydroxyapatite (0001) and (0110) surfaces. *Langmuir* **25**, 5018–5025 (2009)
11. De Leeuw, N.H.: Computer simulations of structures and properties of the biomaterial hydroxyapatite. *J Mater Chem* **20**, 5376 (2010)
12. Almora-Barrios, N., de Leeuw, N.H.: A density functional theory study of the interaction of collagen peptides with hydroxyapatite surfaces. *Langmuir* **26**, 14535–14542 (2010)
13. Chiatti, F., Corno, M., Ugliengo, P.: Stability of the dipolar (001) surface of hydroxyapatite. *J Phys Chem C* **116**, 6108–6114 (2012)
14. Ospina, C.A., Terra, J., Ramirez, A.J., Farina, M., Ellis, D.E., Rossi, A.M.: Experimental evidence and structural modeling of nonstoichiometric (010) surfaces coexisting in hydroxyapatite nano-crystals. *Colloids Surf B Biointerfaces* **89**, 15–22 (2012)
15. Astala, R., Stott, M.: First-principles study of hydroxyapatite surfaces and water adsorption. *Phys Rev B* **78**, 075427 (2008)

16. Slepko, A., Demkov, A.: First principles study of hydroxyapatite surface. *J Chem Phys* **139**, 044714 (2013)
17. Rimola, A., Corno, M., Zicovich-Wilson, C.M., Ugliengo, P.: Ab initio modeling of protein/biomaterial i: competitive adsorption between glycine and water onto hydroxyapatite surfaces. *Phys Chem Chem Phys* **11**, 9005–9007 (2009)
18. Rimola, A., Aschi, M., Orlando, R., Ugliengo, P.: Does adsorption at hydroxyapatite surfaces induce peptide folding? Insights from large-scale B3LYP calculations. *J Am Chem Soc* **134**, 10899–10910 (2012)

Surface Patterning

► [Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies](#)

Surface Plasmon Enhanced Optical Bistability and Optical Switching

Weiqiang Mu and John B. Ketterson

Department of Physics and Astronomy,
Northwestern University, Evanston, IL, USA

Definition

A surface plasmon is a collective oscillation involving the coupled motion of the electrons in a metal and the associated electromagnetic field in small structures and at interfaces. In small structures (of order a wavelength and smaller), the oscillations are localized, whereas at a planar interface separating a metal and a dielectric a propagating mode exists that decays exponentially on both sides of the interface. Plasmons can be excited by fast electrons or, in the right geometry, an incoming electromagnetic wave. Here, optical excitation is considered. An optically bistable system can have two different but stable outputs for the same input signal over some

range. By changing the input beyond some threshold, the output of an optically bistable system can be switched between the two stable states, a feature that can be used to switch the output between two channels.

Optical Bistability and Optical Switching

The development of modern optical communications and optical signal processing requires ultrafast switching of optically encoded information. All-optical switching, which uses light to control light, is a very promising candidate for obtaining high operating speeds. One strategy to achieve all-optical switching is to exploit optical bistability. This phenomenon was first observed in sodium vapor in the 1970s [1]. Soon after that, many materials, gas, liquid, and solid, were shown to produce optical bistability [2–4].

To make a system bistable, both nonlinearity and feedback are required. For an optical system having these two properties, the response function, F , of the system can be defined as

$$F(I_i, I_o) = \frac{I_o}{I_i} \quad (1)$$

where I_i and I_o are respectively the input and output signal. The response function can have various forms, depending on system parameters, which in turn affect the relation between the input and output.

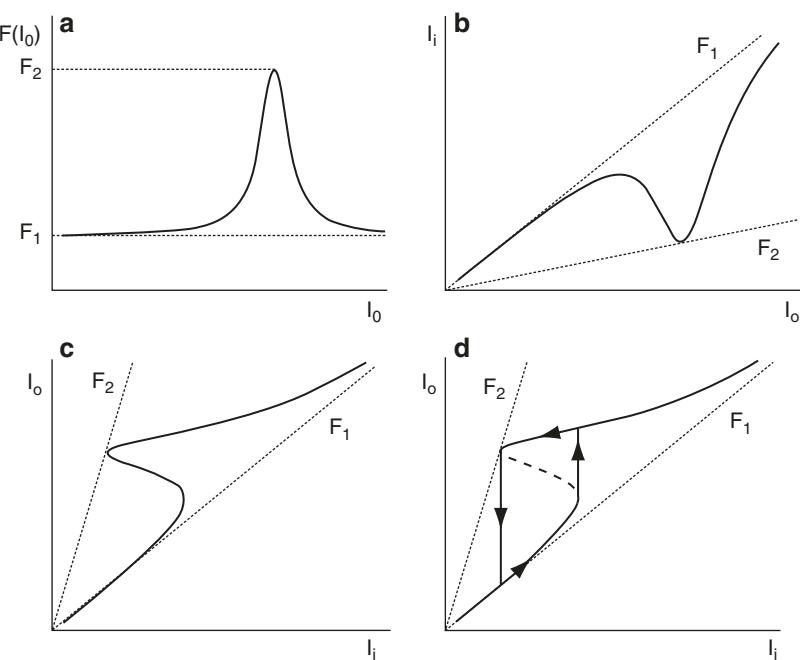
As a simple but important example, consider the effect of the bell-shaped response function shown in Fig. 1a relative to the output signal. Then I_i and I_o are related as

$$I_i = \frac{I_o}{F(I_o)}. \quad (2)$$

Since $F(I_o)$ has a single valued bell-shaped behavior, it is a simple matter to plot I_i using I_o as the independent variable, as shown in Fig. 1b. The corresponding behavior of I_o as a function of I_i is then shown in Fig. 1c. The system is clearly nonlinear, but more importantly for a special range of the inputs the output has three possible

Surface Plasmon Enhanced Optical Bistability and Optical Switching, Fig. 1 (a)

Typical response curve of an optically bistable system; (b) the corresponding I_i relative to I_o behavior; (c) the corresponding I_o relative to I_i behavior. (d) When I_i continuously increases or decreases, I_o shows a hysteresis; the dashed line denotes the unstable region



solutions, as can be seen from Fig. 1c. Those involving a negative slope are unstable and hence the system can be bistable in the region between the two points where the derivative diverges. When continuously varying the input, the output will select one of the stable branches, depending on its history; that is, a kind of hysteresis is achieved as shown in Fig. 1d. As noted, the dashed lines correspond to a region of unstable solutions. The sudden changes and the history dependence of the output can be utilized to make optical switches.

There are many ways to introduce feedback in an optical system. As examples, bistability has been demonstrated in devices based on Fabry–Perot etalons [5], microdisk resonators [6], photonic-crystal cavities [7], etc. The nonlinearities can arise if either refractive index or absorption is intensity dependent. Devices based on the former are said to be dispersive, while the latter are called dissipative. The intensity dependent nonlinearities, which arise from electronic contributions to the third order susceptibility, $\chi^{(3)}$, such as the Kerr effect, can respond in a picosecond or less. While such response times are fast enough for presently envisioned

applications, the magnitude of these nonlinearities is very small; very high laser powers are then required to generate the desired optical bistability.

Surface Plasmons

The Dielectric Constant of Metals

When free electrons near a metal surface are electromagnetically excited, the charge fluctuations are typically confined within a Thomas-Fermi screening length of the metal surface ($\sim 1 \text{ \AA}$). The metals used in plasmonic devices are typically gold, silver, copper, or aluminum. It is common to distinguish two types of surface plasmons: localized surface plasmons (LSP) and propagating surface plasmons (PSP); this primarily depends on the geometry and the distinction is not always sharp. The main difference is that the LSP does not propagate (although, as in the case of a strip resonator, the mode can be considered as counter propagating waves). On the other hand, the PSP transports energy present within the interface. Both originate from the special dielectric properties of metals: a negative real part and a relatively small imaginary part. In the very simplest model

the dielectric constant, ϵ_m , can be described by the Drude-Lorentz form:

$$\epsilon(\omega) = 1 - \frac{\omega_p^2}{\omega^2 + i\omega\gamma} \quad (3)$$

here $\omega_p = 4\pi n e^2 / m$ is the electron plasma frequency, with n and m being the electron density and mass respectively, while γ is the collision rate that governs the damping. The dielectric response is clearly strongly frequency dependent. In visible or near infrared, γ is generally much smaller than ω . In this limit, the real and imaginary parts of ϵ_m can be approximated as

$$\begin{aligned} \epsilon_r(\omega) &= 1 - \frac{\omega_p^2}{\omega^2} \\ \epsilon_i(\omega) &= \frac{\gamma\omega_p^2}{\omega^2} \end{aligned} \quad (4)$$

From Eq. 4, it is clear that when $\omega < \omega_p$, the real part, ϵ_r , is negative.

Equation 3 neglects the host dielectric response, which can be incorporated by replacing the one on the right hand side by a parameter ϵ_∞ ; more complicated forms, involving the addition of Lorentzian-like poles, have been used which give a much better overall representation of the measured optical response [8]. Most metals have a plasma frequency in the ultraviolet or visible. The real part of $\epsilon(\omega)$ for the noble metals gold and silver is negative in the visible and infrared while the imaginary part is small. In addition, they are very stable chemically which makes them the best candidates for plasmonics. Besides metals, some doped semiconductors are good plasmonic materials in the infrared region.

Localized Surface Plasmons

For LSP, the simplest case is a spherical metal particle excited by an optical field $E_0(\omega)$, where the sphere radius, a , is much smaller than the optical wavelength, λ . Using a simple electrostatic model the induced dipole moment is given by

$$\vec{p}(\omega) \propto \frac{\epsilon_m - \epsilon_d}{\epsilon_m - 2\epsilon_d} \vec{E}_0(\omega) \quad (5)$$

Here ϵ_d is the dielectric constant of the medium. For the model based on Eq. 2, it is easy to see that with $\epsilon_d = 1$, the real part of the dominator in Eq. 5 will vanish for $\omega = \omega_p/\sqrt{3}$. Similar behavior happens with more complex models. Because of the relative small value of ϵ_i for Ag and Au, the induced dipole moment can be very large. The scattering from such metal particles is then strong. Clearly, the position of the resonance is determined by the real part of ϵ_m , but the magnitude and bandwidth are determined by the imaginary part of ϵ_m .

Propagating Surface Plasmons

As a surface wave, the propagating surface plasmon can only exist as a transverse-magnetic (TM) wave, as shown the inset of Fig. 2. The electric fields in the two media have the form [9]:

$$\begin{aligned} \vec{E}_d &= (E_{dx}, 0, E_{dz}) \cdot \exp[i \cdot (k_{sp}x - \omega t) - \gamma_d z] \quad (z > 0) \\ \vec{E}_m &= (E_{mx}, 0, E_{mz}) \cdot \exp[i \cdot (k_{sp}x - \omega t) - \gamma_m z] \quad (z > 0) \end{aligned} \quad (6)$$

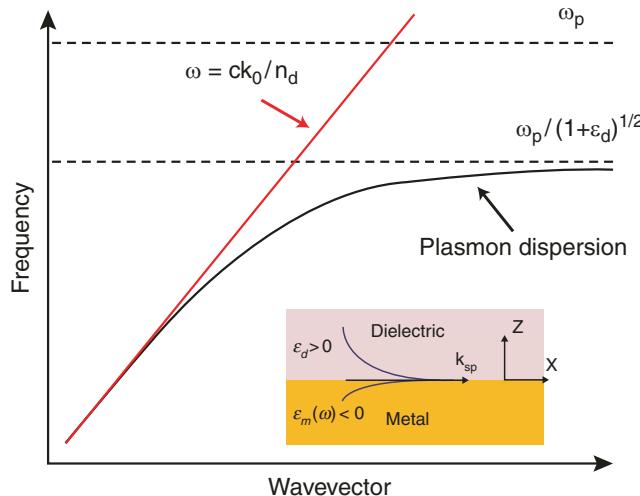
where k_{sp} is the surface plasmon wavevector, and γ_d and γ_m satisfy:

$$\gamma_d = \sqrt{k_{sp}^2 - \epsilon_d k_0^2}; \quad \gamma_m = \sqrt{k_{sp}^2 - \epsilon_m k_0^2} \quad (7)$$

where $k_0 = \omega/c$ is the free space wavevector. In order that the EM wave decays in both sides of the boundary (a requirement for a surface wave), the real part of γ_m and γ_d must be positive. From the conditions that $E_{//}$ and D_{\perp} be continuous, yielding $E_{dx} = E_{mx}$, and $\epsilon_d E_{dz} = \epsilon_m E_{mz}$, the surface plasmon dispersion relation can be derived:

$$k_{sp}(\omega) = k_0 \sqrt{\frac{\epsilon(\omega) \cdot \epsilon_d}{\epsilon(\omega) + \epsilon_d}} \quad (8)$$

The necessary condition for a propagating surface plasmon is that the real part of the quotient under the square root sign in Eq. 8 be positive. This imposes a restriction that $\epsilon_r(\omega) < -\epsilon_d$. The real part of k_{sp} , is the propagation wavevector of the



Surface Plasmon Enhanced Optical Bistability and Optical Switching, Fig. 2 Dispersion relation of propagating surface plasmons based on Eqs. 3 and 8 with $\varepsilon_d = 1$. The red line is the light dispersion in the dielectric. The plasmon dispersion asymptotically

approaches the frequency $\omega_p/(1 + \varepsilon_d)^{1/2}$. The inset is the schematic representation of surface plasmon at the metal-dielectric interface. The wave is TM wave, and the fields decay exponentially in both sides of the interface

surface plasmon. The characteristic plasmon propagation distance is the inverse of the imaginary part of k_{sp} . To achieve long propagation distances, it is necessary to have small ε_i or larger $|\varepsilon_r|$. In this entry, silver is selected as the plasmonic material. Figure 2 shows the dispersion relation for the surface plasmons propagating at the interface between silver and fused silica ($n = 1.4607$ [10]). By fitting the experimental values of the dielectric constant of silver to polynomials, the dielectric constant of silver in visible region can be well represented using the following equation [11]:

$$\begin{aligned} \varepsilon_r &= -255.32 + 198.63\omega - 60.79\omega^2 + 8.38\omega^3 - 0.43\omega^4 \\ \varepsilon_i &= 83.26 - 132.79\omega + 90.47\omega^3 + 6.66\omega^4 \\ &\quad - 0.71\omega^5 + 0.03\omega^6 \end{aligned} \quad (9)$$

where $\omega = 2\pi \cdot c/\lambda \cdot 10^{-15} S^{-1}$.

Because the PSP wavevector is larger than that in the dielectric, the mode cannot be directly excited from the dielectric. Two popular methods to excite PSP are grating coupling and prism coupling. The grating coupling [12–14] utilizes a diffraction grating to provide the extra

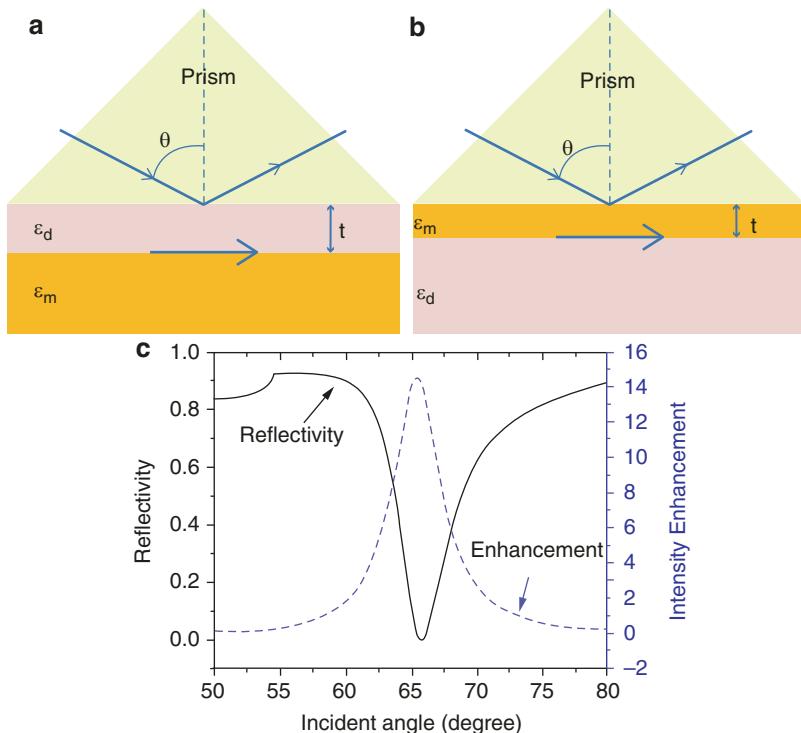
contribution to the wavevector of the photons propagating in the dielectric side to match the real part of k_{sp} (this is called an umklapp process in solid state physics). The grating coupling occurs when

$$k_{//} = k_{sp}^r + mK_g m = 0, \pm 1, \pm 2 \dots \quad (10)$$

where $k_{//}$ is the in-plane wavevector of the incoming wave, k_{sp}^r is the real part of the surface plasmon wavevector, and K_g is the grating wavevector. The coupling efficiency depends on the parameters of the grating such as the ditch depth and ditch shape.

The two common prism coupling schemes, the Otto [15] and the Kretschmann [16] geometries (named after their originators), are shown in Fig. 3a, b. For both configurations, the coupling prisms must have a higher refractive index than the dielectric at the metal interface. By correctly choosing the incident beam parameters, such as the incident angle or the wavelength, the in-plane wavevector $k_{//}$ of TM mode incoming beams can be matched to k_{sp}^r for the surface plasmon between metal and the dielectric. If the coupling layer between the prism surface and the plasmon

Surface Plasmon Enhanced Optical Bistability and Optical Switching, Fig. 3 (a) The Otto geometry and (b) the Kretschmann configuration for coupling to a propagating surface plasmon mode with a TM optical wave by total internal reflection. (c) The reflectivity and the intensity enhancement of the surface plasmon with Kretschmann configuration



interface is not too thick, the evanescent field from the prism surface will excite the metal surface. In the Otto configuration, the coupling from the prism to the metal film is through a layer with dielectric constant smaller than that of the prism, as in Fig. 3a. In the Kretschmann configuration, the coupling occurs through the metal film itself.

When light is coupled into the PSP with prism coupling, the reflectivity, which occurs in a regime that would otherwise correspond to total internal reflection (TIR), will have a dip; this is called *attenuated total internal reflection* (ATR). This results from part of the energy of the incoming beam has been coupled into the PSP wave, and partially from dissipation by the ohmic losses in the metal. For both methods, the coupling efficiency, or the dip depth, depends on the thickness of the coupling layer. If the coupling layer is too thick, only a small portion of the incoming beam is coupled into surface plasmon; this is referred to as undercoupling. On the other hand, if the coupling layer is too thin, a large portion of the excited surface plasmon will be coupled back into the reflection beam and the buildup of the

wave is limited; this is called overcoupling. To get the optimum coupling, which occurs where the reflected beam intensity vanishes and all the energy goes into the mode, the coupling layer has to have just the right thickness.

The reflectivity and the intensity enhancement associated with the PSP can be modeled with the multilayer Fresnel reflection method. The intensity enhancement is defined as the ratio of the intensity at the metal-dielectric boundary to the incoming intensity in the prism. Figure 3c shows simulation results with the Kretschmann configuration. In the simulation, the light wavelength was fixed at 532 nm, corresponding to the second harmonic frequency of a Nd:YAG laser, and the incident angle is scanned from 50° to 80°. The refractive indices of prism (SF11) and dielectric material (fused silica) are 1.7948 and 1.4607 respectively. The silver thickness has been optimized as 44.5 nm to obtain the best coupling.

Surface Plasmon-Enhanced Optical Bistability
Surface plasmons can be exploited to significantly enhance optical bistability. First, as shown in the

simulation results of Fig. 3c, the local optical field at the interface is greatly enhanced. This reduces the needed laser driving power. Secondly, the high sensitivity of surface plasmons to the change in the refractive index of the adjacent dielectric makes for a fast and sensitive switch. Surface plasmons (in many geometries) can then be utilized to produce optical bistability. In what follows, several approaches are described to achieve plasmon-based bistability.

The first approach that exploits plasmon induced optical bistability utilizes the Kretschmann configuration [17, 18]. If the dielectric material adjacent to the silver surface supporting the plasmons is replaced by a nonlinear Kerr medium, then the refractive index will depend on the local optical intensity. The refractive index n_k of Kerr material is written as (When the surface plasmon is excited, the actual optical intensity in the Kerr medium is exponentially decaying. In the simulations presented here, the value of n_k is taken to be that calculated with the intensity *at the interface* as a simplified model and neglect its change in the Kerr medium, as in Ref. [17]. In Ref. [18], it is shown that if the intensity distribution is considered, the bistability shape will be same, but the required optical intensity will be higher than the value predicted by this simple model)

$$n_k = n_0 + n_2 \cdot I = n_0 + n_2 \cdot A \cdot I_{in}; \quad (11)$$

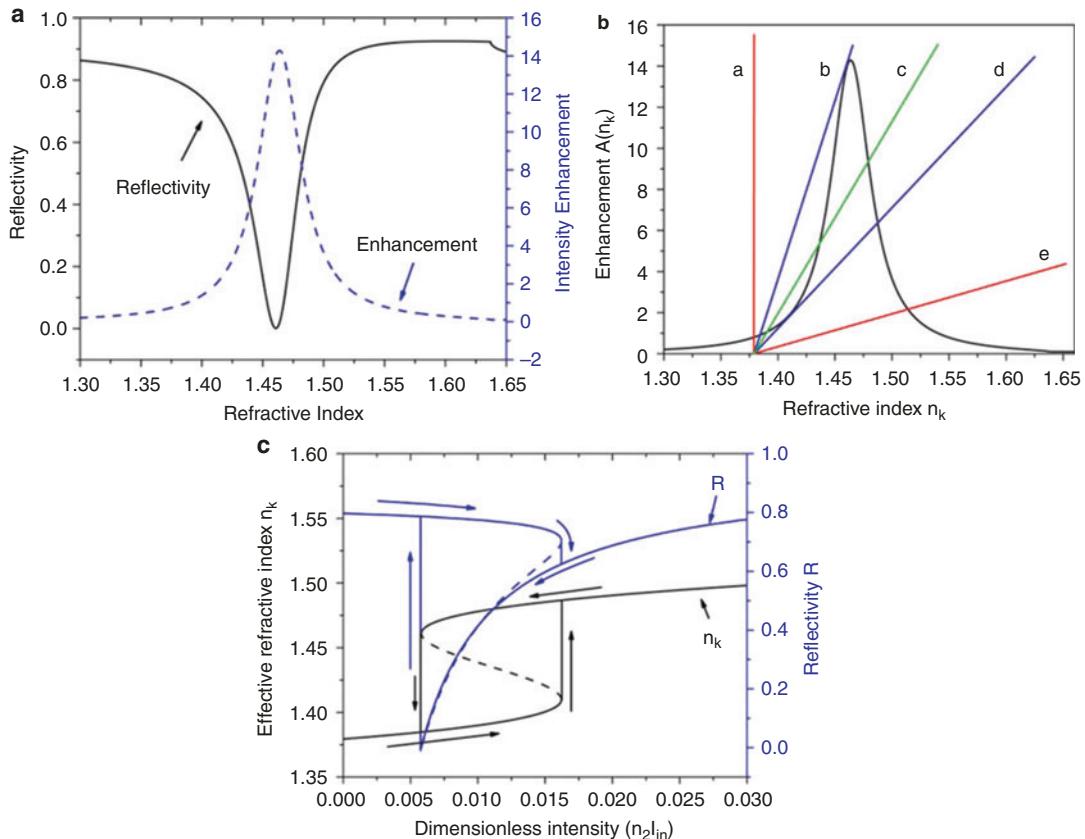
here, n_0 and n_2 are the linear and second order nonlinear refractive index of the Kerr material, I the light intensity in the Kerr medium, I_{in} the input optical intensity, and A is the intensity enhancement due to the surface plasmon wave. The Kerr medium examined here is a mixture of nitrobenzene and ethanol. The values of n_0 and n_2 for nitrobenzene are 1.553 and 2.766×10^{-20} (m/V)² respectively [19]; the refractive index of ethanol is 1.362. By mixing the two liquids in the right ratio, the linear refractive index n_2 can be adjusted to lie between 1.362 and 1.553. The effective n_2 will be assumed to be proportional to the volume percentage of the nitrobenzene.

Using the same geometry as for the simulation of Fig. 3c but with the incident angle fixed at

65.81 ° (which corresponds to the best coupling in Fig. 3c), and scanning the refractive index of the dielectric medium from 1.3 to 1.65, the corresponding intensity enhancement factor A (n_k) is plotted as the dashed line in Fig. 4a. With the calculated $A(n_k)$, the values of n_k based on Eq. 11 will exhibit bistability relative to the incident intensity, I_{in} . This is graphically depicted in Fig. 4b [20]. Equation 11 is first rewritten as:

$$A(n_k) = \frac{n_k - n_0}{n_2 I_{in}} = \frac{n_k - n_0}{I_d} \quad (12)$$

here I_d is defined as the dimensionless input intensity. The left side of Eq. 12 $A(n_k)$ is plotted as the black curve in Fig. 4b. With fixed value of n_0 , the right side of Eq. 12 is straight line, the slope of which is proportional to the reciprocal of I_{in} (or I_d). The values of n_k are then determined by the intersections of the black curve and the straight line. For the simulation shown in Fig. 4b, the value of n_0 is selected to be 1.38. On the other hand, the resonance, which is the peak position of curve $A(n_k)$, occurs at $n_k = 1.46$. The difference between these two values corresponds to three times the Δn associated with the full width at half maximum (FWHM) of the resonance. Note all of the lines, resulting from different I_{in} , intersect at $n_k = n_0$, as seen in Fig. 4b. As I_{in} increases from zero, the slope of the lines continuously decreases, from $a \rightarrow b \rightarrow c \rightarrow d \rightarrow e \rightarrow$. For the regions with very low and very high input intensity (the region between lines a and b and the region beyond line d), there is only one solution for n_k . Line b and line d correspond to critical points, where there are two solutions to the equation. Between line b and line d, there are three possible values of n_k for the same input, line c being an example. When there are multiple solutions, the physical system selects one, depending on its history. Figure 4c shows the effective index of refraction (n_k) and the reflectivity (R) as a function of the dimensionless input intensity; note the bistability in both n_k and R is clearly presented. The dashed parts of the line are the unstable solutions.



Surface Plasmon Enhanced Optical Bistability and Optical Switching, Fig. 4 (a) The reflectivity and intensity enhancement relative to the refractive index of the dielectric in the Kretschmann configuration. (b) Graphic solutions for n_k with different optical input intensities. The black curve is the intensity enhancement $A(n_k)$ relative to n_k ; (c) Solutions for n_k and the corresponding values of reflectivity for different input intensities. By continuously changing the input intensity, both n_k and the reflectivity exhibit bistability

the different straight lines represent different input intensities; the intersections between the *straight lines* and the *enhancement curve* give the solution for n_k . (c) Solutions for n_k and the corresponding values of reflectivity for different input intensities. By continuously changing the input intensity, both n_k and the reflectivity exhibit bistability

Long-Range Surface Plasmon-Based Optical Bistability

For most of the nonlinear materials, the Kerr coefficient n_2 in Eq. 11 is very small. As in Fig. 4c, to achieve the bistability, the required dimensionless intensity needs to be at least 0.017, corresponding to an input optical intensity on the order of 10^{12} W/cm^2 . Therefore, to have a strong optical bistability, it is better to have a sharper resonance, and with it a higher enhancement of the local field. Long-range surface plasmon polaritons (LRSPP), a mode that exists under special conditions, possesses these two characteristics. Of course, a

narrower resonance comes at the cost of a longer response time.

If a thin metal film is sandwiched between dielectric materials having the same optical properties, surface plasmon waves will exist on *both* sides of the metal film. When the metal film is thin enough, the two surface plasmons will interact with each other and split into two modes: a symmetric mode and an antisymmetric mode relative to the in-plane electric field at the film center. The electric fields on the two surfaces are in phase for symmetric mode and 180° out of phase for the antisymmetric mode. In a manner similar to that used to obtain the dispersion relation for SPP, and

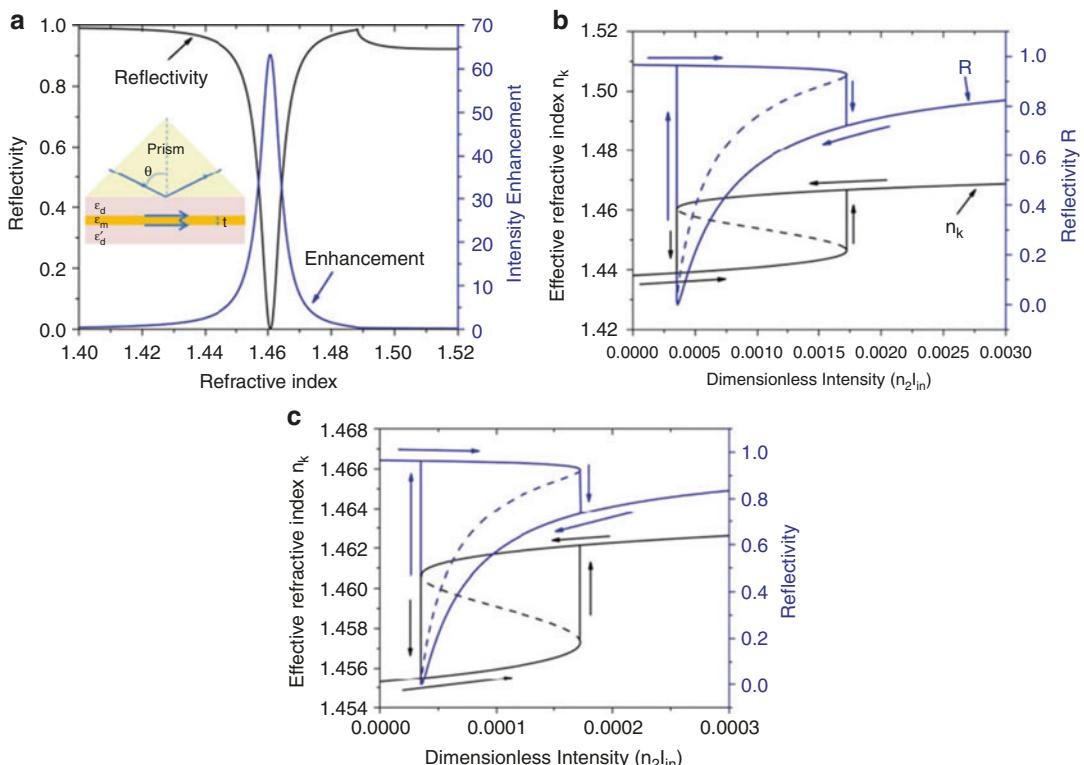
after imposing the boundary conditions, the wavevectors for these two surface plasmon modes are as [21]:

$$\begin{aligned}\tanh\left(\frac{\gamma_m t}{2}\right) &= \frac{\epsilon_m \gamma_d}{\epsilon_d \gamma_m} \text{ (anti-symmetric mode)} \\ \tanh\left(\frac{\gamma_m t}{2}\right) &= \frac{\epsilon_d \gamma_m}{\epsilon_m \gamma_d} \text{ (symmetric mode)}\end{aligned}\quad (13)$$

where t is the thickness of the metal layer, and $\gamma_i = \sqrt{k_{sp}^2 - \epsilon_i k_{0i}^2}$. For the symmetric mode, the optical fields constructively interfere inside the metal, which, in turn, results in an increase in the ohmic damping. Therefore, the plasmon lifetime (and with it the decay length) is relatively small; hence it is called a short-range surface plasmon polariton (SRSPP). For the antisymmetric

mode, the electric fields from the two surfaces destructively interfere inside the metal, thereby decreasing the ohmic losses. The plasmon lifetime (and the decay length) then become much longer; this mode is called the long-range surface plasmon polariton (LRSPP). As in Eq. 13, the wavevectors of these surface plasmons depend on the metal thickness t . In the thin film limit, the damping of LRSPP is approximately proportional to t^2 . Therefore, in principle, by continuously decreasing the metal thickness, we should have very small losses and hence very large local optical intensities.

In order to be able to compare with the optical bistability induced by conventional PSP shown in Fig. 4, the same calculations were repeated for the LRSPPs. The inset of Fig. 5a shows the geometry of the excitation of LRSPP using the ATR coupling method. When the silver thickness is 20 nm,



Surface Plasmon Enhanced Optical Bistability and Optical Switching, Fig. 5 (a) The reflectivity and intensity enhancement of the LRSPP when it is excited via the ATR method. The *inset* shows the geometry used to excite the LRSPP. (b) The bistability of refractive index of the Kerr material and the reflectivity relative to the dimensionless intensity when silver film thickness is 20 nm. (c) The bistability of refractive index of the Kerr material and the reflectivity relative to the dimensionless intensity when silver film thickness is 10 nm

the Kerr material and the reflectivity relative to the dimensionless intensity for the case where the silver film thickness is 20 nm. (c) The bistability of refractive index of the Kerr material and the reflectivity relative to the dimensionless intensity when silver film thickness is 10 nm

perfect coupling (the absence of reflection) happens when the coupling layer, here a fused silica layer between the silver and the prism, is 555 nm, and the incident angle is 56.02°. Then the last layer of the dielectric is replaced with a material for which the refractive index ranges between 1.40 and 1.52. The reflectivity and the intensity enhancement are the black and blue curves respectively. Clearly, compared with Fig. 4a, the peak intensity enhancement resulting from the LRSPP is much larger, while the width of the bell-shape enhancement is much narrower. To directly compare the enhancement between the LRSPP and conventional PSP, the difference between n_0 and the resonance of intensity enhancement curve $A(n_k)$ in Fig. 5b is still set to be three time the Δn associated with the full width at half maximum (FWHM) of the resonance of $A(n_k)$. Figure 5b shows the resulting bistability of the refractive index n_k for the Kerr medium and the total reflectivity. Clearly, the peak laser intensity needed to produce the optical bistability here is much smaller than the value needed in Fig. 4c, decreasing by one order of magnitude (from 0.017 to 0.0015). By further decreasing the silver film thickness, the needed peak intensity to produce bistability will decreases even more; this is demonstrated in Fig. 5c for the case where the silver film is 10 nm. Here, the optimized coupling layer thickness is 965 nm and coupling angle 54.87°.

Conclusions

The resonant behavior exhibited by conventional propagating surface plasmon polaritons and especially the long-range propagating plasmon polaritons can be exploited to enhance the optical intensity in an adjacent dielectric media. When combined with a nonlinear Kerr medium, one can produce optical bistability and with it the ability to form an optical switch. The effect can be manifested as a change in the reflectivity or a phase delay from plasmons excited in an ATR geometry [22], a change in the scattering intensity from localized surface plasmons [23], a shift in the transmission or reflection coefficient in a grating-coupled surface plasmon geometry [24], and so

on. To lower the pumping threshold for bistability, geometries leading to a sharper resonance, and with it a larger intensity enhancement, must be exploited.

Cross-References

- Active Plasmonic Devices
- Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures
- Surface Plasmon-Polariton-Based Detectors

References

1. Gibbs, H.M., McCall, S.L., Venkatesan, T.N.C.: Differential gain and bistability using a Sodium-filled Fabry-Perot interferometer. *Phys. Rev. Lett.* **36**, 1135–1138 (1976)
2. Abraham, E., Smith, S.D.: Optical bistability and related devices. *Rep. Prog. Phys.* **45**, 815–885 (1982)
3. Gibbs, H.M.: Optical Bistability: Controlling Light with Light. Academic, Orlando (1985)
4. Gibbs, H.M., McCall, S.L., Venkatesan, T.N., Gossard, A.C., Passner, A., Wiegmann, W.: Optical bistability in semiconductors. *Appl. Phys. Lett.* **36**, 451–453 (1979)
5. Felber, F.S., Marburger, J.H.: Theory of nonresonant multistable optical devices. *Appl. Phys. Lett.* **28**, 731–733 (1976)
6. Borselli, M., Johnson, T.J., Painter, O.: Beyond the Rayleigh scattering limit in high-Q silicon microdisks, theory and experiment. *Opt. Express* **13**, 1515–1530 (2005)
7. Notomi, M., Shinya, A., Mitsugi, S., Kira, G., Kuramochi, E., Tanabe, T.: Optical bistable switching action of Si high-Q photonic-crystal nanocavities. *Opt. Express* **13**, 2678–2687 (2005)
8. Sukharev, M., Sievert, P.R., Seideman, T., Ketterson, J.B.: Perfect coupling of light to surface plasmons with ultra-narrow linewidths. *J. Chem. Phys.* **131**, 034708 (2009)
9. Rather, H.: Surface Plasmons on Smooth and Rough Surfaces and on Gratings. Springer, Berlin/New York (1988)
10. Palik, E.D.: Handbook of Optical Constants of Solids. Academic, New York (1985)
11. Chen, Z., Hooper, I.R., Sambles, J.R.: Strongly coupled surface plasmons on thin shallow metallic gratings. *Phys. Rev. B* **77**, 161405 (2008)
12. Mu, W., Buchholz, D.B., Sukharev, M., Jang, J.I., Chang, R.P.H., Ketterson, J.B.: One-dimensional long-range plasmonic-photonic structures. *Opt. Lett.* **35**, 550–552 (2010)

13. Chen, Y.J., Koteles, E.S., Seymour, R.J., Sonek, G.J., Ballantyne, J.M.: Solid State Commun. **46**, 95 (1983)
14. Gruhlke, R.W., Holland, W.R., Hall, D.G.: Optical emission from coupled surface plasmons. Opt. Lett. **12**, 364–366 (1987)
15. Otto, A.: Excitation of nonradiative surface plasma waves in silver by method of frustrated total reflection. Z. Phys. **216**, 398–410 (1968)
16. Kretschmann, E.: Z. Phys. **241**, 313 (1971)
17. Wysin, G.M., Simon, H.J., Deck, R.T.: Optical bistability with surface plasmon. Opt. Lett. **6**, 30–32 (1981)
18. Hickernell, R.K., Sarid, D.: Optical bistability using prism-coupled, long-range surface plasmons. J. Opt. Soc. Am. B **3**, 1059–1069 (1986)
19. Boyd, R.W.: Nonlinear Optics. Academic, Boston (2003)
20. Saleh, B.E.A., Teich, M.C.: Fundamentals of Photonics. Wiley, New York (1991)
21. Sarid, D.: Long-range surface-plasma waves on very thin metal films. Phys. Rev. Lett. **47**, 1927–1931 (1981)
22. Nazvanov, V.F., Kovalenko, D.I.: Phase optical bistability in structures with surface plasmons. Tech. Phys. Lett. **24**, 650–651 (1998)
23. Leung, K.M.: Optical bistability in the scattering and absorption of light from nonlinear microparticles. Phys. Rev. A **33**, 2461–2464 (1986)
24. Wurtz, G.A., Pollard, R., Zayats, A.V.: Optical bistability in nonlinear surface-plasmon polaritonic crystals. Phys. Rev. Lett. **97**, 057402 (2006)

Surface Plasmon Nanophotonics

► Optical Properties of Metal Nanoparticles

Surface Plasmon Polariton-Enabled High-Performance Organic Optoelectronic Devices

Jing Feng and Hong-Bo Sun
 State Key Laboratory on Integrated Optoelectronics, College of Electronic Science and Engineering, Jilin University, Changchun, China

Introduction

In organic optoelectronic devices, such as organic light-emitting devices (OLEDs) and organic solar

cells (OSCs), high efficiency is one of the key issues for their applications, and much effort has been devoted to developing novel materials and device structures [1, 2]. It is well known that the majority of the generated light is trapped in OLEDs. Around 80 % of internally generated light is trapped in the form of waveguide (WG) modes in organic and indium tin oxide (ITO) anode layers and in surface plasmon polariton (SPP) modes associated with the metallic electrode/organic interface in OLEDs [3, 4]. In organic solar cells (OSCs), the active layer is generally less than 100 nm due to the short exciton diffusion length, which limits the efficiency of incident light absorption. A thicker active layer offers higher light absorption; however, it comes at the expense of lowered exciton harvesting [5, 6]. So far, there is still the greatest scope for significant improvements about efficiency in OLEDs and OSCs.

Microstructure with wavelength- to subwavelength-scale periodicity has played an important role in optical and optoelectronic devices, particularly in optical fibers, distributed feedback lasers, LEDs, and solar cells, through manipulating the generation and propagation of photons in materials. Introducing a microstructure onto the metallic electrode is especially crucial for recovering the power lost to the associated SPP modes in OLEDs by providing an additional momentum to couple the SPP modes [3, 7]. Plasmonic nanostructures have been introduced into solar cells for highly efficient light harvesting. Two types of plasmonic resonances, surface plasmonic resonances (SPRs) [8, 9] and localized plasmonic resonances (LPRs) [10, 11], can be used for enhancing light absorption in an organic solar cell without increasing the thickness of its active layers. This entry concerns recent progresses in fabrication of periodic microstructure in organic optoelectronic devices and their effects on improving the device performance. The SPP modes that were generally lost in conventional OLEDs have been successfully recovered by employing the periodic corrugations, and a much enhanced light extraction has been observed. The introduction of the microstructures into OSCs is effective in relieving the mismatch

between optical absorption length and charge transport scale, and enhanced light absorption in thinner organic films has been achieved.

OLEDs Integrated with Microstructures

Solving Efficiency-Stability Tradeoff in TOLEDs by Employing Periodically Corrugated Metallic Cathode

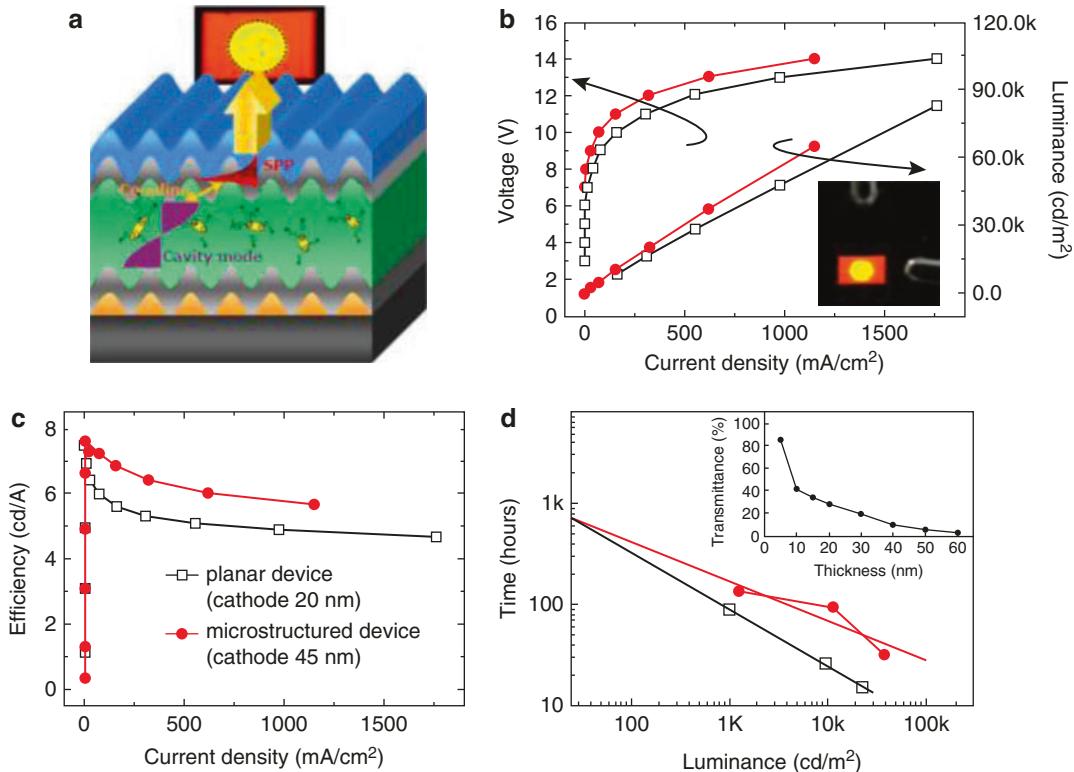
In conventional top-emitting organic light-emitting devices (TOLEDs), there exists a fundamental tradeoff between device stability and efficiency, which is really a challenge for their further applications. The semitransparent top cathode of a conventional TOLED is usually composed of multiple functional layers to achieve both optical transmission and effective electron injection [12]. An ultrathin layer of a reactive, low work function metal (~ 1 nm) is used for electron injection, and a noble metal with high transparency is used to reduce the sheet resistance of the composite cathode and also as a protective layer for the underlying reactive metal and organic layers, for example, bilayer cathodes of Al/Ag, Ca/Ag, etc. Ag is a widely used protective layer because of its low optical absorption and high conductivity. The Ag layer should be less than 20 nm to ensure a high transparency; however, more pinholes are formed in the Ag film because of poor film continuity for such a thin film, which results in easy diffusion of water and oxygen from air into the devices and accelerates the degradation of the devices [13, 14]. Thicker Ag layers offer higher film continuity; however, the higher device stability comes at the expense of lower efficiency because of the decreased optical transmittance.

Employing a periodic corrugation in the metal cathode of the TOLEDs provides a possibility to optimally solve the tradeoff [15]. SPP-mediated emission from an OLED incorporating a periodic wavelength-scale microstructure has been observed [16, 17]. In the case of a thin metal film, SPP excitation on the opposite interface can be supported, and they can couple to each other by the familiar grating-coupling mechanism [18, 19]. This results in a much enhanced light

transmission through the classically opaque metal film [20] and permits a relatively thick metal film of ~ 50 nm [21], which is much more effective in protecting the devices from exposure to atmosphere. In the case of the TOLEDs, the reflective anode and semitransparent cathode are parallel to each other and form a microcavity. The cavity length is fixed by the distance between the two parallel electrodes and defines its resonant wavelength. In this case, the cross coupling occurs between the SPP modes associated with the top surface of the cathode and the microcavity modes within the device, instead of the two SPPs, which also results in an enhanced light transmission. Taking into account that the SPP resonance can be tuned by adjusting the period of the corrugated metal film or the refractive index of the dielectric on the metal surface to coincide with the microcavity modes, the light transmission can be enhanced at a desired wavelength. Therefore, employing a periodically corrugated cathode opens a possible avenue to release the efficiency-stability tradeoff that persists in TOLEDs. Figure 1a demonstrates that the introduction of a periodic corrugation into TOLEDs to realize a corrugated metal cathode is effective in relieving the tradeoff between device stability and efficiency, through the cross coupling of the SPPs associated with the Ag cathode surface and the microcavity modes within the TOLEDs [4]. The thickness of the Ag cathode for the corrugated TOLEDs was increased from 20 to 45 nm, and both the device lifetime and efficiency are significantly improved, which has been examined by performing current density (J)-voltage (V)-luminance (L) and lifetime measurements of the electrically pumped TOLEDs as shown in Fig. 1b–d.

Surface Plasmon Polariton-Mediated Red Emission from OLEDs Based on Metallic Electrodes Integrated with Dual-Periodic Corrugation

Using metallic anode in OLEDs, for example, thin metal film, metal grid, or metal nanowires with high optical transmission and electrical conductivity as a direct replacement for ITO, has a potential to recover the power lost to the WG modes in ITO [22, 23]. OLEDs with metallic films as both



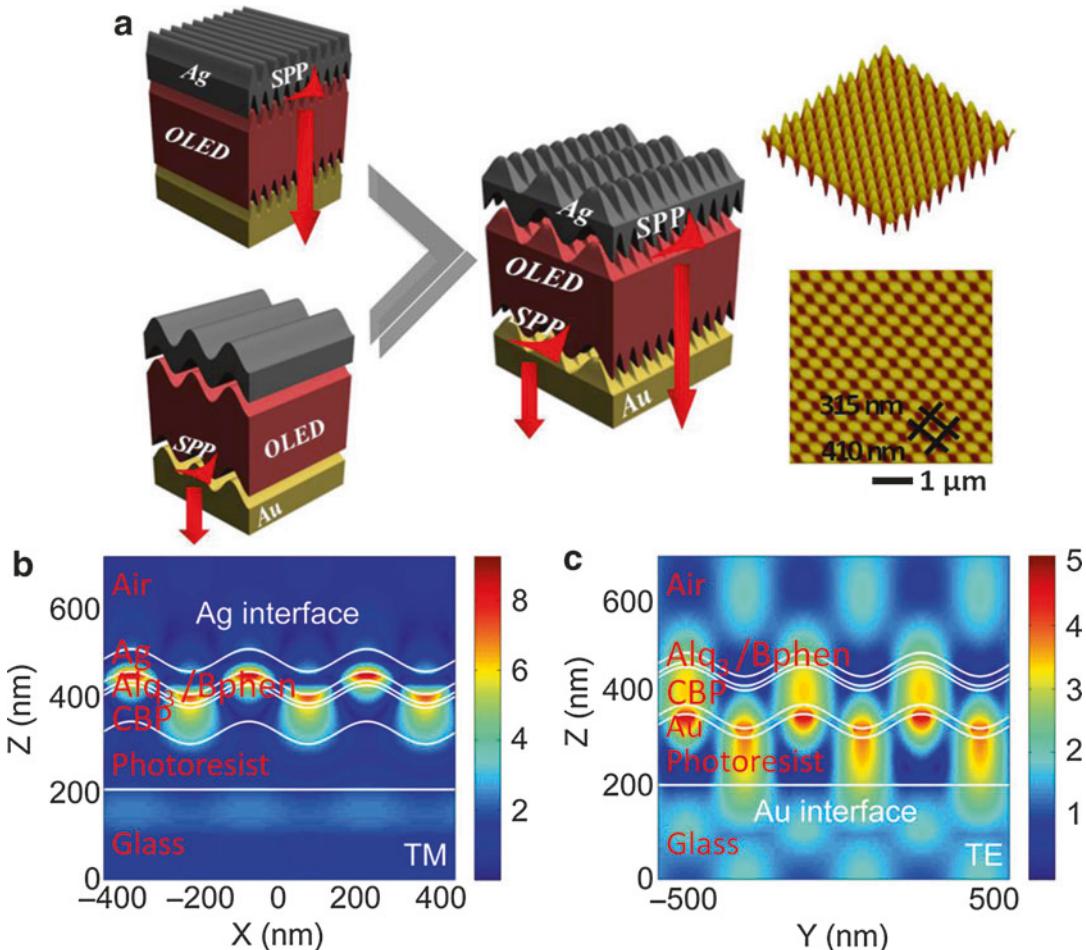
Surface Plasmon Polariton-Enabled High-Performance Organic Optoelectronic Devices, **Fig. 1** (a) Schematic cross section of the red TOLED with periodic microstructure. (b-d) EL performance of the corrugated and planar TOLEDs. Current density-voltage-luminance (c), current density efficiency (d) characteristics, and half-luminance lifetime (e), respectively.

Inset in (d) shows the photograph of the operating corrugated and planar OLEDs at the same substrate and under same driving voltage. *Inset in (d)* shows the transmittance of the Ag film with various thicknesses from 5 to 60 nm. Reproduced with permission from Ref. [4]. Copyright 2012, Wiley-VCH

top cathode and bottom anode by employing semitransparent Au thin film as anode have shown superior device performance [3, 24]. In this case, the light trapped in the SPP modes existed at cathode/organic interface and anode/organic interface become the main power lost. Introducing a wavelength-scale microstructure onto the metallic electrode surface has exhibited its remarkable effect for excitation and outcoupling of the SPP modes by providing an additional momentum to couple the SPP modes into light [4, 25]. A periodic microstructure is suitable for specific narrow range of wavelength by satisfying the Bragg scattering condition. SPP resonant wavelength would be different for the cathode and anode interface due to the different

metal materials used, and the excitation and outcoupling of the SPP modes in the desired wavelength region can be realized only at one of the electrode interfaces. Therefore, recovering the power lost to SPPs at both cathode and anode interfaces, simultaneously, is an important issue to further improve the efficiency of the OLEDs.

Efficient excitation and outcoupling of the SPP modes associated with both top and bottom electrode/organic interfaces have been realized by introducing a two-dimensional (2-D) grating with dual-periodic corrugation into the OLEDs (Fig. 2a) [26]. The 2-D grating consists two sets of corrugations with different periods. The SPP resonance at the cathode/organic and anode/organic interfaces can be tuned to the same



Surface Plasmon Polariton-Enabled High-Performance Organic Optoelectronic Devices, Fig. 2 (a) Schematic of excitation and outcoupling of the SPP modes associated with both cathode/organic and anode/organic interfaces in OLEDs by engaging dual-periodic corrugation and AFM images of surface morphologies of photoresist with 2-D dual-periodic corrugation.

Distributions of the magnetic field intensity across the dual-periodic sample with the normal incident light at the wavelength of 605 nm with TM polarization associated with Ag/organic interface (b) and TE polarization associated with Au/organic interface (c). Reproduced with permission from Ref. [26]. Copyright 2014 Nature Publishing Group

wavelength to coincide with the electroluminescent (EL) peak of the OLEDs by adjusting the appropriate periods of the two sets of corrugations. As a result, the light trapped in the SPP modes associated with the two electrode/organic interfaces are both efficiently extracted from the OLEDs, respectively. In-house generated finite-difference time-domain (FDTD) code is applied to simulate the magnetic field intensity distributions across the corrugated devices as shown in Figs. 2b and 3c. The numerical simulations

confirm the excitation and outcoupling of the SPP modes from both top and bottom metal/organic interfaces in the dual-periodic corrugated devices.

Broadband Light Extraction from White OLEDs by Employing Corrugated Metallic Electrodes with Dual Periodicity

Broadband extraction is important for the efficient outcoupling of the trapped photons associated with substrate mode, SPP, and WG modes,

especially for that trapped in the white organic light-emitting devices (WOLEDs) whose spectra covering the whole visible wavelength. It can be easily realized for broadband extraction of the substrate modes by attaching a microlens array on the outside of the substrate. While in the case of the SPP and WG modes inside of the OLEDs, structure modification has to be introduced inside the device structure, and broadband extraction is difficult to be obtained. Wavelength-scale periodic microstructures introduced into the OLEDs are suitable for specific narrow range of wavelength by satisfying the Bragg scattering condition and applicable only for monochromatic OLEDs [4, 27]. Other microstructures, such as spontaneously formed buckles or defective hexagonal close-packed arrays, have been used to effectively enhance light extraction for OLEDs [28], while their effect on broadband light extraction in WOLEDs has not yet been examined. So far, broadband light extraction from WOLEDs is still a challenge, which is an obstacle for its applications in both display and lighting.

OLEDs with two metallic electrodes by employing metallic film with high optical transmission and electrical conductivity to replace ITO as anode could eliminate the power lost to the WG modes in ITO [3, 29]. In this case, light trapped in the SPP modes are the main power lost, and therefore, highly efficient light extraction could be expected by broadband excitation and outcoupling of the SPP modes in the WOLEDs. Broadband excitation and outcoupling of the SPP modes in the WOLEDs have been realized by introducing a 2-D grating with dual-periodic corrugation into the WOLEDs (Fig. 3a) [24]. The 2-D grating which consists two sets of corrugations with different periods can broaden the SPP resonance compared to that of the monoperiodic grating (Fig. 3b). The blue and orange emissions are both efficiently extracted from the WOLEDs based on the two complementary color strategies by adjusting appropriate periods of the dual-periodic corrugation. Both experimental and numerical results support the validity of the broadband light extraction, and a 37 % enhancement in current efficiency compared

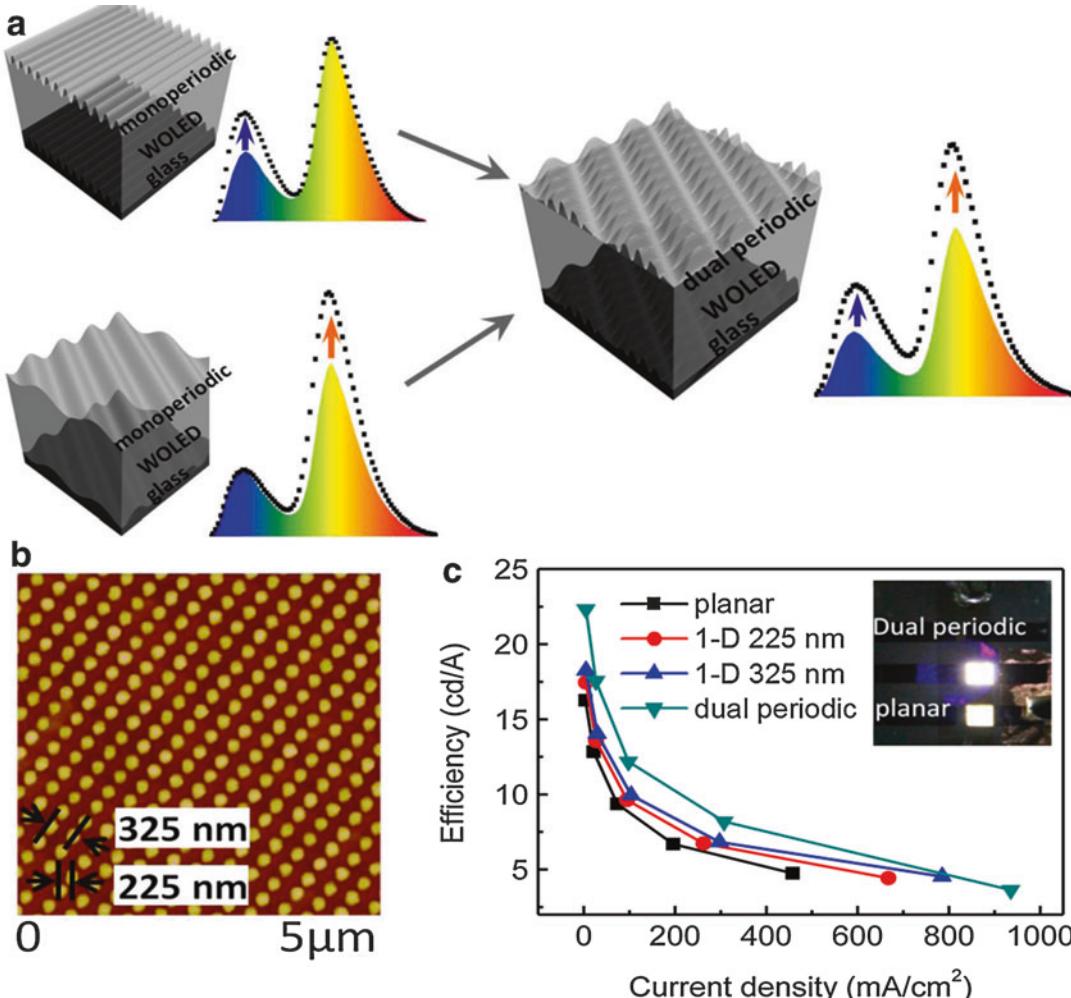
to those of the conventional planar devices has been obtained (Fig. 3c) [24].

OSCs Integrated with Microstructures

Surface Plasmon-Enhanced Absorption in OSCs by Employing a Periodically Corrugated Metallic Electrode

OSCs based on conjugated polymers and small molecules are an attractive alternative to Si-based solar cells due to their advantages of low cost, light weight, simple fabrication process, and flexibility [30]. However, the power conversion efficiency (PCE) of the OSCs is still low for commercial applications. The tradeoff between the efficiency of photon absorption and exciton harvesting is one of the main limitation factors. The active layer is generally less than 100 nm due to the short exciton diffusion length, which limits the efficiency of incident light absorption [31]. A thicker active layer offers higher light absorption; however, it comes at the expense of lowered exciton harvesting. Therefore, an exploration of device design to improve light absorption without increasing the physical thickness of the photovoltaic absorber layer is required.

A periodically nanopatterned metal film forms a particularly interesting class, since their periodicity can be tuned to adjust the SPP resonance and the metal film itself can be used as metallic electrode in the devices. Up to 50 % increment of the absorption in the active layers of the OSCs has been predicted by previous theory analysis [32]. Therefore, using the periodically nanopatterned metal to excite SPP may provide a possible way to release the tradeoff between photon absorption and exciton harvesting efficiency. Such a corrugated metallic electrode with wavelength-scale periodicity was integrated into the OSC structure to excite the SPP at the metal/organic interface with tunable resonance [33]. The SPP resonance has been tuned by changing the grating period, so that a coincidence between the SPP resonance and the absorption wavelength region of the absorber layer can be obtained to realize a maximum absorption enhancement. The short-circuit current (J_{sc}) and the PCE were



Surface Plasmon Polariton-Enabled High-Performance Organic Optoelectronic Devices,

Fig. 3 (a) Schematic of the broadband light extraction by using the dual-periodic corrugation. (b) AFM images of surface morphologies of photoresist with 2-D dual-periodic corrugation. (c) EL performance of the corrugated

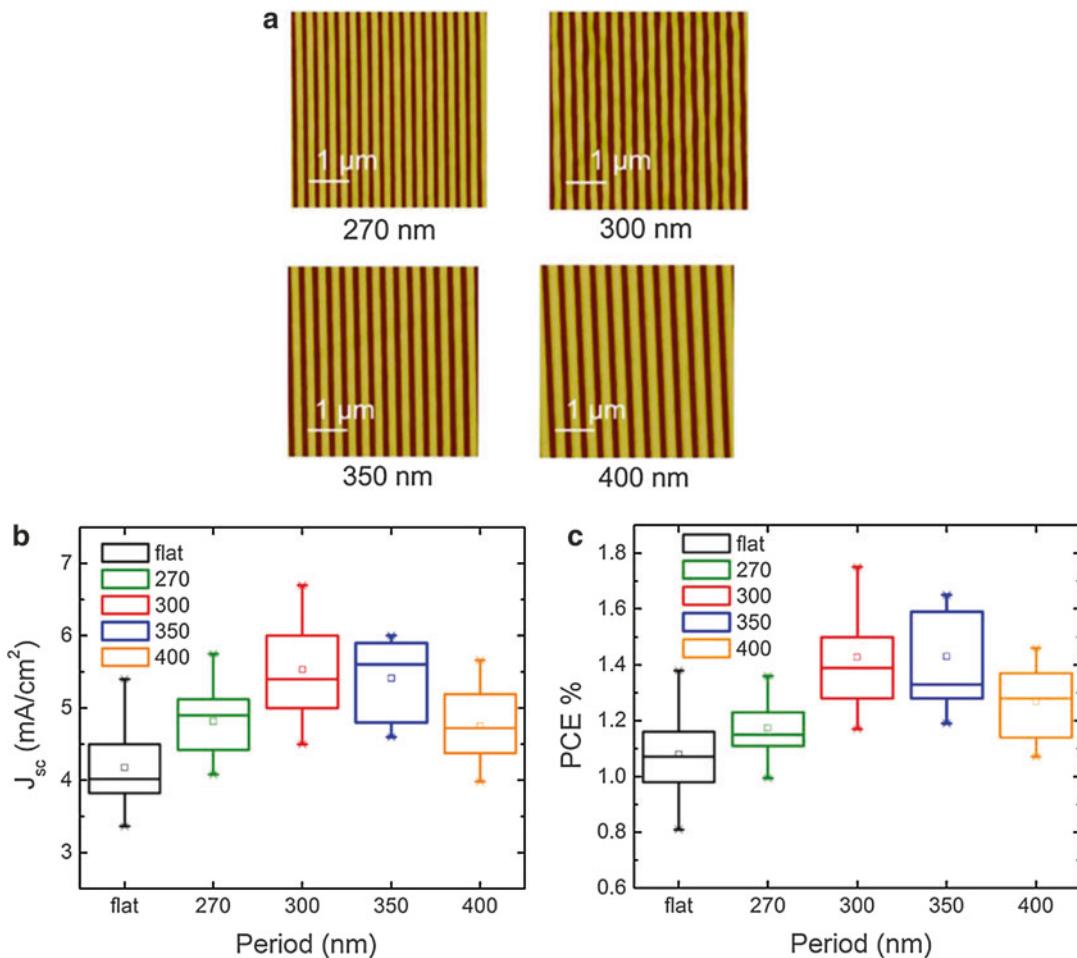
and planar WOLEDs. The *inset* in (c) shows the photograph of the operating dual-periodic and planar WOLEDs at the same substrate and under the same driving voltage. Reproduced with permission from Ref. [24]. Copyright 2013, Wiley-VCH

significantly improved for the corrugated OSCs with appropriate period (Fig. 4).

Effective and Tunable Light Trapping in Bulk Heterojunction Organic Solar Cells by Employing Au-Ag Alloy Nanoparticles

Various light trapping approaches have been explored to solve the conflict about the thickness of the active layer by increasing the absorption [33, 34]. The incorporation of metallic nanoparticles (NPs) into the OSC structures supports

the amplification of the electric field near the particle surface through localized SPR (LSPR) and could result in enhanced light absorption owing to the near-field coupling. Au and Ag NPs have been studied in great detail in the OSCs due to their relative strong scattering efficiency, broad LSPR absorption band in the visible range, and chemical stability [35, 36]. Mixing of the Au and Ag NPs has also been investigated in OSCs, which result in larger efficiency enhancement due to a dual resonance enhancement [37].



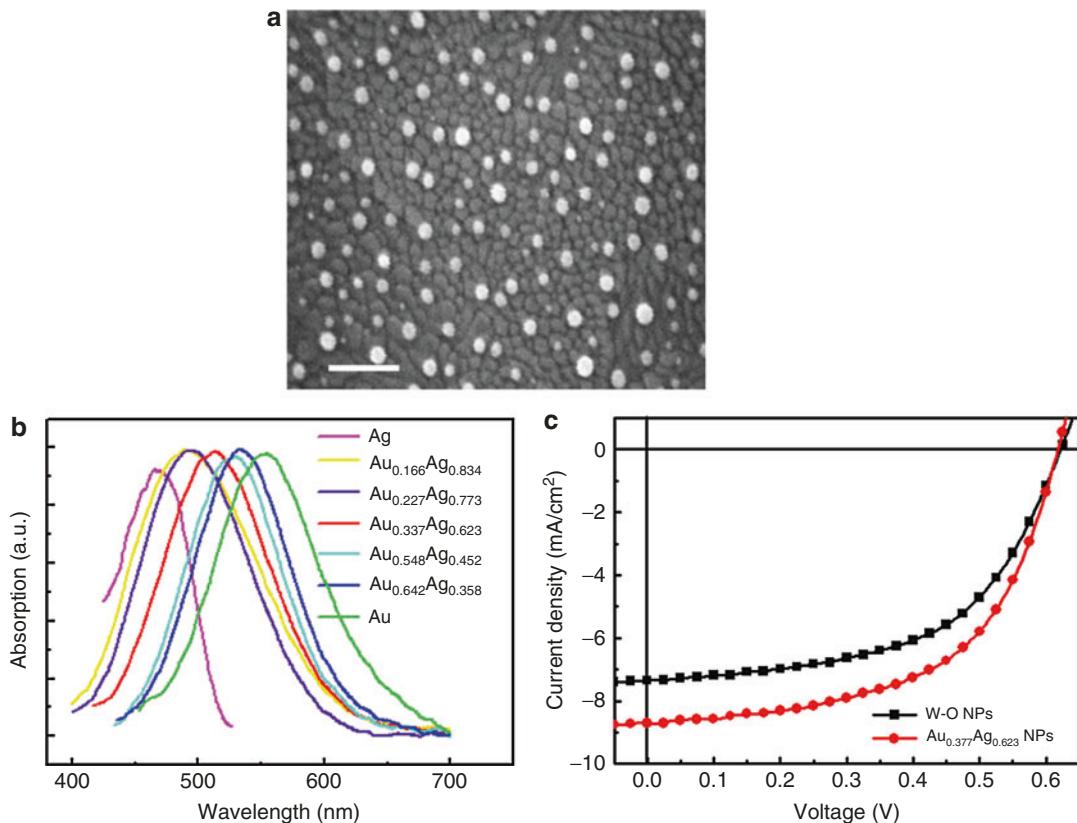
Surface Plasmon Polariton-Enabled High-Performance Organic Optoelectronic Devices, Fig. 4 (a) AFM images of surface morphology of periodic corrugation with period of 270, 300, 350, and 400 nm. Multiple device statistical performance analysis of the

short-circuit current (b) and the power conversion efficiency (c) produced in planar device and the corrugated devices with various periods. Reproduced with permission from Ref. [33]. Copyright 2012 AIP Publishing LLC

Compared to monometallic NPs and physical mixing of the Au and Ag NPs, the Au-Ag alloy NPs have more advantages due to the potential to combine the best of both metals in terms of plasmonic properties and exhibit unique electronic, optical, and catalytic properties that distinct from those of the corresponding monometallic particles [38, 39]. Moreover, Au and Ag easily form alloys for different compositions with very little surface segregation due to their similar lattice constants [40]. Different from the monometallic NPs, for which both the shape and size of the metallic NPs are key factors

determining the LSPR wavelength [41], the LSPR can be tuned easily to the desired spectral range by varying the molar ratio of the Au-Ag alloy NPs. Therefore, the use of the Au-Ag alloy NPs in the OSCs offers a great potential for effective and tunable light trapping.

The thermal annealing method has been employed to fabricate the Au-Ag alloy NPs with tunable molar ratio, and its effect on light trapping in the OSCs has been investigated. The Au-Ag alloy NPs were obtained by a co-evaporation of Au and Ag onto the ITO substrate and followed by the thermally annealing process, so that the molar



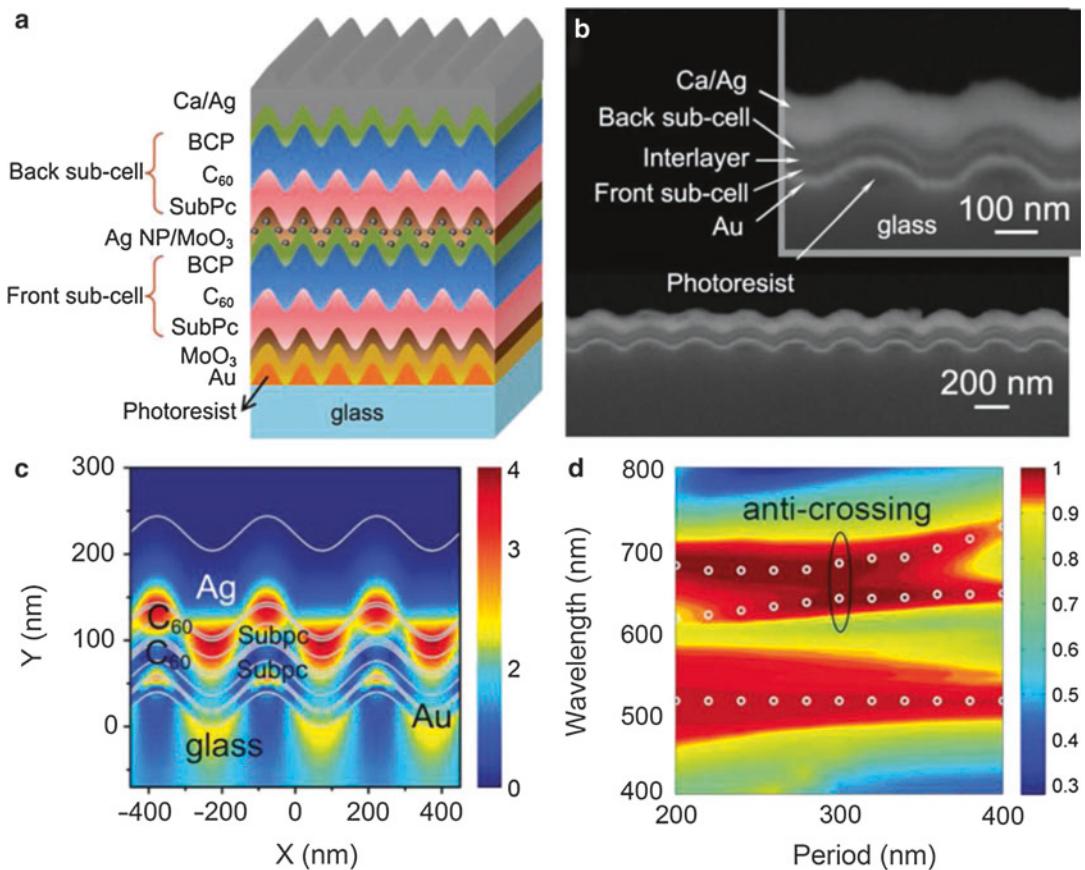
Surface Plasmon Polariton-Enabled High-Performance Organic Optoelectronic Devices, Fig. 5 (a) SEM images of the Au-Ag alloy NPs formed by vacuum annealing on the ITO substrates. (b) UV-vis absorption spectra of Au and Ag monometallic NPs and

Au-Ag alloy NPs with various molar ratios. (c) J-V characteristics of the OSCs with the $\text{Au}_{0.377}\text{Ag}_{0.623}$ alloy NPs and without the NPs. Reproduced with permission from Ref. [42]. Copyright 2014 AIP Publishing LLC

ratio of the alloy NPs can be easily and precisely tuned by controlling the evaporation rate of the Au and Ag individually (Fig. 5a) [42]. Light absorption enhancement in the active layer has been obtained by strong LSPR excited at the absorption wavelength region of the absorber by the alloy NPs with an appropriate molar ratio (Fig. 5b). Meanwhile, light scattering from the NPs elongates the optical path length which contributes to the light absorption enhancement at a shorter wavelength. The short-circuit photocurrent density (J_{sc}) of the alloy NP-based OSCs is improved from 7.37 to 8.74 mA/cm², and the efficiency of 3.03 % is obtained, which corresponds to an enhancement factor of 19 % (Fig. 5c).

Matching Photocurrents of Sub-cells in Double-Junction Organic Solar Cells via Coupling Between Surface Plasmon Polaritons and Microcavity Modes

Tandem configurations, in which two or more single cells are stacked in series, have been accepted as a successful and universal strategy to increase the V_{oc} of the OSCs [43]. A record of up to 7 V for six-junction solar cells has been reported [44]. However, the photocurrent of the tandem cells is usually lower than that of the single-junction cells, because less light is absorbed by the back sub-cell compared to the front one in a typical tandem device, which creates an imbalance in the photocurrent generated by the



Surface Plasmon Polariton-Enabled High-Performance Organic Optoelectronic Devices, Fig. 6 (a) Schematic structure of the corrugated double-junction solar cells. (b) SEM cross section of the corrugated device. (c) Distribution of the magnetic field intensity in the corrugated double-junction cells with period of

300 nm at the wavelength of 650 nm. (d) Dispersion relationship for the wavelength versus the period of the corrugation of the corrugated double-junction cells. Reproduced with permission from Ref. [34]. Copyright 2013, Wiley-VCH

back and front sub-cells and limits the PCE of the overall device. Appropriate choice of materials with different bandgaps can overcome this problem to a certain degree [45]. However, this approach is difficult to apply to multi-junction organic solar cells, especially for small molecule-based cells, due to the limited choice of the small molecule donor materials that have significantly different absorption profiles. Careful design of the optical field and photovoltaic contributions from the sub-cells is another strategy to optimize the tandem cell. Surface plasmon polariton (SPP)-induced field enhancement has been believed to be a highly attractive solution to enhance optical

absorption in an organic solar cell without increasing the thickness of its active layers [33, 46].

SPP resonance has been employed in the small molecule-based double-junction organic solar cells with two identical sub-cells to enhance the optical absorption of the back sub-cell by employing periodical corrugation in the metallic cathode (Fig. 6) [34]. Both the SPP and microcavity modes are supported by the corrugated device structure, and the SPP resonance is tuned by tuning the period of the corrugation, so that an anti-crossing between the SPP and microcavity modes within the device is realized. This anti-crossing plays an important role in

enhancing the absorption of the back sub-cell and thereafter achieving a balanced photocurrent of front and back sub-cells, which results in an improved PCE for the double-junction device. Both simulation and experimental results support the effect of the anti-crossing behavior on the absorption enhancement. For the double-junction devices with this periodical corrugation, 10.4 % enhancement in the photocurrent and 11.3 % enhancement in the PCE have been achieved.

Conclusions

Microstructures have been introduced into optoelectronic devices, and enhanced performances have been demonstrated. The introduction of periodic corrugations has allowed the observation of the emission originating from the SPP modes, which are usually trapped within planar OLEDs. By employing a periodic microstructure in TOLEDs, a much enhanced light transmission through a thick cathode has been observed through the grating-induced cross coupling between the SPPs associated with top interface of the cathode and microcavity modes within the device cavity. Dual-periodic corrugation has been introduced into WOLED metallic electrodes to realize broadband light extraction. Improved efficiency of OSCs by employing a periodically corrugated metallic electrode has been demonstrated. Light absorption enhancement in the active layer has been obtained by strong LSPR excited at the absorption wavelength region of the absorber by the alloy NPs with an appropriate molar ratio. By employing a periodic wavelength-scale corrugation into the device structure, the optical absorption of the back sub-cell in the double-junction organic solar cells has been enhanced by strong anti-crossing coupling between the microcavity and propagating SPP modes at the Ag/back sub-cell interface.

Cross-References

- [Light Emitting Diodes](#)
- [Organic Photovoltaics: Basic Concepts and Device Physics](#)

References

1. Park, S.H., et al.: Bulk heterojunction solar cells with internal quantum efficiency approaching 100%. *Nat. Photon.* **3**, 297–303 (2009)
2. D'Andrade, B.W., Forrest, S.R.: White organic light-emitting devices for solid-state lighting. *Adv. Mater.* **16**, 1585–1595 (2004)
3. Bi, Y.G., et al.: Enhanced efficiency of organic light-emitting devices with metallic electrodes by integrating periodically corrugated structure. *Appl. Phys. Lett.* **100**, 053304 (2012)
4. Jin, Y., et al.: Solving efficiency-stability tradeoff in top-emitting organic light-emitting devices by employing periodically corrugated metallic cathode. *Adv. Mater.* **24**, 1187–1191 (2012)
5. Liu, L., Stanchina, W.E., Li, G.: Effects of semiconducting and metallic single-walled carbon nanotubes on performance of bulk heterojunction organic solar cells. *Appl. Phys. Lett.* **94**, 233309 (2009)
6. Brabec, C.J., Sariciftci, N.S., Hummelen, J.C.: Plastic solar cells. *Adv. Funct. Mater.* **11**, 15–26 (2001)
7. Nalwa, K.S., Park, J.M., Ho, K.M., Chaudhary, S.: On realizing higher efficiency polymer solar cells using a textured substrate platform. *Adv. Mater.* **23**, 112 (2011)
8. Baba, A., Wakatsuki, K., Shinbo, K., Kato, K., Kaneko, F.: Increased short-circuit current in grating-coupled surface plasmon resonance field-enhanced dye-sensitized solar cells. *J. Mater. Chem.* **21**, 16436–16441 (2011)
9. Sha, W.E.I., Choy, W.C.H., Chew, W.C.: A comprehensive study for the plasmonic thin-film solar cell with periodic structure. *Opt. Express* **18**, 5993–6007 (2010)
10. Yang, J., et al.: Plasmonic polymer tandem solar cell. *ACS Nano* **5**, 6210–6217 (2011)
11. Wang, D.H., et al.: Enhanced power conversion efficiency in PCDTBT/PC70BM bulk heterojunction photovoltaic devices with embedded silver nanoparticle clusters. *Adv. Energy Mater.* **1**, 766–770 (2011)
12. Hung, L.S., Tang, C.W., Mason, M.G., Raychaudhuri, P., Madathil, J.: Application of an ultrathin LiF/Al bilayer in organic surface-emitting diodes. *Appl. Phys. Lett.* **78**, 544–546 (2001)
13. Lim, S.F., Ke, L., Wang, W., Chua, S.J.: Correlation between dark spot growth and pinhole size in organic light-emitting diodes. *Appl. Phys. Lett.* **78**, 2116–2118 (2001)
14. Kolosov, D., et al.: Direct observation of structural changes in organic light emitting devices during degradation. *J. Appl. Phys.* **90**, 3242–3247 (2001)
15. Atwater, H.A., Polman, A.: Plasmonics for improved photovoltaic devices. *Nat. Mater.* **9**, 205–213 (2010)
16. Hobson, P.A., Wedge, S., Wasey, J.A.E., Sage, I., Barnes, W.L.: Surface plasmon mediated emission from organic light-emitting diodes. *Adv. Mater.* **14**, 1393–1396 (2002)

17. Lupton, J.M., Matterson, B.J., Samuel, I.D.W., Jory, M.J., Barnes, W.L.: Bragg scattering from periodically microstructured light emitting diodes. *Appl. Phys. Lett.* **77**, 3340–3342 (2000)
18. Ebbesen, T.W., Lezec, H.J., Ghaemi, H.F., Thio, T., Wolff, P.A.: Extraordinary optical transmission through sub-wavelength hole arrays. *Nature* **391**, 667–669 (1998)
19. Hooper, I.R., Sambles, J.R.: Coupled surface plasmon polaritons on thin metal slabs corrugated on both surfaces. *Phys. Rev. B* **70**, 045421 (2004)
20. Gruhlke, R.W., Holland, W.R., Hall, D.G.: Surface plasmon cross coupling in molecular fluorescence near a corrugated thin metal film. *Phys. Rev. Lett.* **56**, 2838–2841 (1986)
21. Wedge, S., Hooper, I.R., Sage, I., Barnes, W.L.: Light emission through a corrugated metal film: the role of cross-coupled surface plasmon polaritons. *Phys. Rev. B* **69**, 245418 (2004)
22. Kuang, P., et al.: A new architecture for transparent electrodes: relieving the trade-off between electrical conductivity and optical transmittance. *Adv. Mater.* **23**, 2469–2473 (2011)
23. Chiappe, D., Toma, A., de Mongeot, F.B.: Transparent plasmonic nanowire electrodes via self-organised ion beam nanopatterning. *Small* **9**, 913–919 (2013)
24. Bi, Y.G., et al.: Broadband light extraction from white organic light-emitting devices by employing corrugated metallic electrodes with dual periodicity. *Adv. Mater.* **25**, 6969–6974 (2013)
25. Yates, C.J., Samuel, I.D.W., Burn, P.L., Wedge, S., Barnes, W.L.: Surface plasmon-polariton mediated emission from phosphorescent dendrimer light-emitting diodes. *Appl. Phys. Lett.* **88**, 161105 (2006)
26. Bi, Y.G., et al.: Surface plasmon-polariton mediated red emission from organic light-emitting devices based on metallic electrodes integrated with dual-periodic corrugation. *Sci. Rep.* **4**, 7108 (2014)
27. Bai, Y., et al.: Outcoupling of trapped optical modes in organic light-emitting devices with one-step fabricated periodic corrugation by laser ablation. *Org. Electron.* **12**, 1927–1935 (2011)
28. Koo, W.H., et al.: Light extraction of organic light emitting diodes by defective hexagonal-close-packed array. *Adv. Funct. Mater.* **22**, 3454–3459 (2012)
29. Helander, M.G., et al.: Oxidized gold thin films: an effective material for high-performance flexible organic optoelectronics. *Adv. Mater.* **22**, 2037–2040 (2010)
30. Peumans, P., Forrest, S.R.: Very-high-efficiency double-heterostructure copper phthalocyanine/C60 photovoltaic cells. *Appl. Phys. Lett.* **79**, 126–128 (2001)
31. Agrawal, M., Peumans, P.: Broadband optical absorption enhancement through coherent light trapping in thin-film photovoltaic cells. *Opt. Express* **16**, 5385–5396 (2008)
32. Min, C., et al.: Enhancement of optical absorption in thin-film organic solar cells through the excitation of plasmonic modes in metallic gratings. *Appl. Phys. Lett.* **96**, 133302 (2010)
33. Jin, Y., et al.: Surface-plasmon enhanced absorption in organic solar cells by employing a periodically corrugated metallic electrode. *Appl. Phys. Lett.* **101**, 163303 (2012)
34. Jin, Y., et al.: Matching photocurrents of sub-cells in double-junction organic solar cells via coupling between surface plasmon polaritons and microcavity modes. *Adv. Opt. Mater.* **1**, 809–813 (2013)
35. Rand, B.P., Peumans, P., Forrest, S.R.: Long-range absorption enhancement in organic tandem thin-film solar cells containing silver nanoclusters. *J. Appl. Phys.* **96**, 7519–7526 (2004)
36. Tong, S.W., et al.: Improvement in the hole collection of polymer solar cells by utilizing gold nanoparticle buffer layer. *Chem. Phys. Lett.* **453**, 73–76 (2008)
37. Lu, L., Luo, Z., Xu, T., Yu, L.: Cooperative plasmonic effect of Ag and Au nanoparticles on enhancing performance of polymer solar cells. *Nano Lett.* **13**, 59–64 (2013)
38. Qingbo, Z., Jim Yang, L., Jun, Y., Chris, B., Jixuan, Z.: Size and composition tunable Ag–Au alloy nanoparticles by replacement reactions. *Nanotechnology* **18**, 245605 (2007)
39. Kelly, K.L., Coronado, E., Zhao, L.L., Schatz, G.C.: The optical properties of metal nanoparticles: the influence of size, shape, and dielectric environment. *J. Phys. Chem. B* **107**, 668–677 (2003)
40. Li, Z.Y., et al.: Structures and optical properties of 4–5 nm bimetallic AgAu nanoparticles. *Faraday Discuss.* **138**, 363–373 (2008)
41. Catchpole, K.R., Polman, A.: Design principles for particle plasmon enhanced solar cells. *Appl. Phys. Lett.* **93**, 191113 (2008)
42. Xu, M., et al.: Effective and tunable light trapping in bulk heterojunction organic solar cells by employing Au–Ag alloy nanoparticles. *Appl. Phys. Lett.* **105**, 153303 (2014)
43. Cheyns, D., Rand, B.P., Heremans, P.: Organic tandem solar cells with complementary absorbing layers and a high open-circuit voltage. *Appl. Phys. Lett.* **97**, 033301 (2010)
44. Sullivan, P., et al.: Ultra-high voltage multijunction organic solar cells for low-power electronic applications. *Adv. Energy Mater.* **3**, 239–244 (2013)
45. Dou, L., et al.: Tandem polymer solar cells featuring a spectrally matched low-bandgap polymer. *Nat. Photon.* **6**, 180–185 (2012)
46. Duche, D., et al.: Improving light absorption in organic solar cells by plasmonic contribution. *Solar Energy Mater. Solar Cells* **93**, 1377–1382 (2009)

Surface Plasmon Resonance

► Quantum System Near Metallic Particle

Surface Plasmon-Coupled Emission

- ▶ [Plasmonic Amplification for Fluorescence Bioassays Utilizing Propagating Surface Plasmons](#)

Surface Plasmon-Polariton Photodetectors

- ▶ [Surface Plasmon-Polariton-Based Detectors](#)

Surface Plasmon-Polariton-Based Detectors

Pierre Berini
 School of Information Technology and Engineering (SITE), University of Ottawa, Ottawa, ON, Canada

Synonyms

- [Surface plasmon-polariton photodetectors](#)

Definition

A surface plasmon polariton is a transverse-magnetic optical surface wave propagating along the interface of a metal and a dielectric; it is a coupled excitation formed from electromagnetic fields coupled to a charge density wave in the metal. A surface plasmon-polariton detector is a device capable of detecting surface plasmons or involving surface plasmons in the detection process.

Notation

An $e^{+j\omega t}$ time-harmonic dependence is assumed. The relative permittivity is denoted ϵ_r , and is written for a metal in terms of real and imaginary

parts as $\epsilon_{r,m} = -\epsilon_R - j\epsilon_I$. \mathbf{k} denotes a wavevector (in general). $k_0 = 2\pi/\lambda_0 = \omega/c_0$ is the wavenumber of plane waves in vacuum, λ_0 the wavelength in vacuum, c_0 the speed of light in vacuum, and $\omega = 2\pi\nu$ is the angular frequency.

Single-Interface Surface Plasmon-Polariton

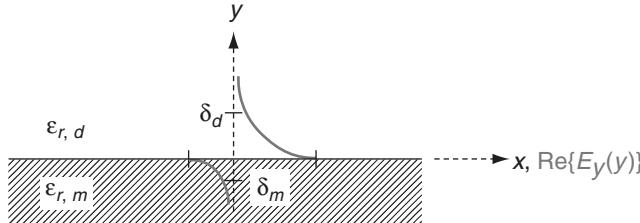
A surface plasmon-polariton (SPP) waveguide is a metallo-dielectric structure along which SPP modes are guided. The simplest structure supporting SPPs is an interface between an optically semi-infinite dielectric and an optically semi-infinite metal [1], as sketched in Fig. 1 (the relative permittivity of the metal is denoted $\epsilon_{r,m}$ and that of the dielectric $\epsilon_{r,d}$) This structure is termed the single-interface.

Highly conductive metals and good dielectrics are used to implement SPP waveguides. Ag is often preferred because it has the lowest optical loss among the metals over a broad wavelength range, but it is reactive, so care must be taken during fabrication and use in order to avoid degradation (which can occur if Ag is exposed to, e.g., air or water). Au is also a good choice given its chemical stability and its good optical performance. A good metal for supporting SPPs normally satisfies $\epsilon_R \gg \epsilon_I$.

Metals are dispersive at optical wavelengths. Away from interband transitions, the Drude model for the permittivity captures the dispersive character of metals [1]:

$$\begin{aligned}\epsilon_{r,m} &= -\epsilon_R - j\epsilon_I \\ &= 1 - \frac{\omega_p^2}{\omega^2 + 1/\tau^2} - j \frac{\omega_p^2/\tau}{\omega(\omega^2 + 1/\tau^2)}\end{aligned}\quad (1)$$

In the above, ω_p is the plasma frequency and τ the relaxation time. The “Drude region” corresponds to that portion of the electromagnetic spectrum where Eq. 1 holds. For many metals, this region spans the range from visible wavelengths to the infrared. The metal approaches a perfect electric conductor as the wavelength increases through the infrared and beyond.

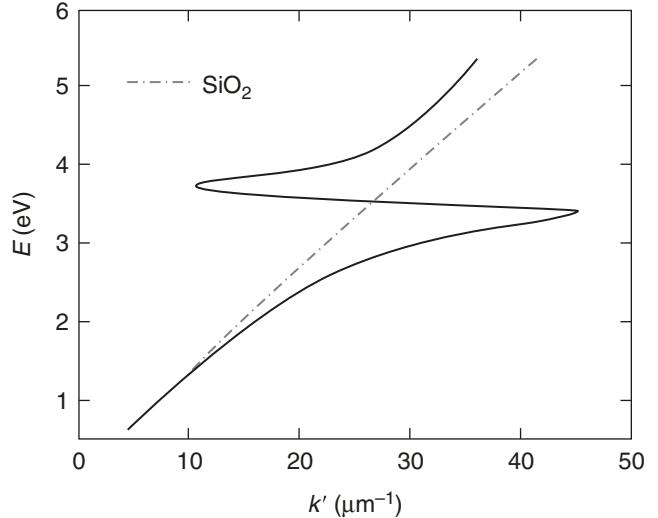


Surface Plasmon-Polariton-Based Detectors,
Fig. 1 Single-interface surface plasmon-polariton waveguide; the distribution of the main transverse electric field

component (E_y) of the SPP is sketched on the structure as the *thick gray curve*

Surface Plasmon-Polariton-Based Detectors,

Fig. 2 Dispersion of the SPP along a Ag/SiO₂ interface. The *light line* in SiO₂ is plotted as the *dash-dot curve*



The single-interface (Fig. 1) supports one purely bound (nonradiative) SPP mode. This mode is transverse magnetic (TM) and may propagate at any angle in the x - z plane (e.g., along $+z$). The SPP fields (E_x , E_z , and H_x) are confined along the y direction, peaking at the interface and decaying exponentially into both media. The distribution of the main transverse electric field component (E_y) of the SPP is sketched on the structure of Fig. 1 as the thick gray curve. The field penetration depth in the metal δ_m is much smaller than the field penetration depth in the dielectric δ_d . Confinement of the SPP arises because the metal and dielectric have $\text{Re}\{\epsilon_r\}$ of opposite sign at the wavelength of operation (indeed over a large wavelength range).

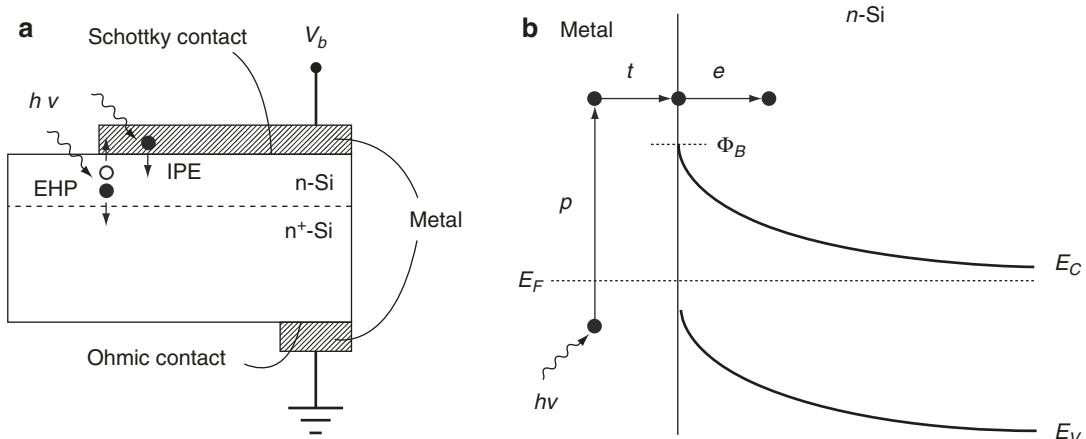
The wavenumber of the single-interface SPP is [1]:

$$k^{\text{SPP}} = k_0 \left(\frac{\epsilon_{r,m} \epsilon_{r,d}}{\epsilon_{r,m} + \epsilon_{r,d}} \right)^{\frac{1}{2}} \quad (2)$$

For a lossless dielectric cladding ($\text{Im}\{\epsilon_{r,d}\} = 0$), the above simplifies to the following approximate expressions for the real (phase) and imaginary (attenuation) parts of k^{SPP} [1]:

$$\begin{aligned} k' &\cong k_0 \left(\frac{\epsilon_R \epsilon_{r,d}}{\epsilon_R - \epsilon_{r,d}} \right)^{\frac{1}{2}} \text{ and} \\ k'' &\cong k_0 \frac{\epsilon_I}{2\epsilon_R^2} \left(\frac{\epsilon_{r,d} \epsilon_R}{\epsilon_R - \epsilon_{r,d}} \right)^{\frac{1}{2}} \end{aligned} \quad (3)$$

Figure 2 plots the dispersion curve of the SPP ($E = hv$ in eV versus k' , h is Planck's constant) on a Ag/SiO₂ interface [2]; the light line in SiO₂ is also plotted as the dash-dot curve. As the frequency decreases (increasing wavelength), the metal approaches a PEC, the confinement of the



Surface Plasmon-Polariton-Based Detectors,
Fig. 3 (a) Schottky diode on n-Si illuminated by light of photon energy $h\nu$. Reverse biasing ($V_b < 0$) is assumed. Electrons are depicted by filled circles and holes by unfilled ones. (b) Energy band diagram of a Schottky

contact on n-Si and the three-step internal photoemission process: p photoexcitation, t transport, e emission. E_C and E_V are the conduction and valence band edges, respectively, E_F is the Fermi level, and Φ_B is the Schottky barrier height

SPP decreases, and its dispersion curve merges with the light line (SPPs are not supported at the interface between a smooth PEC and a semi-infinite dielectric). As the frequency increases, the SPP approaches an “energy asymptote,” readily observed in Fig. 2 near $E \sim 3.4$ eV ($\lambda_0 \sim 360$ nm) where k' diverges. The group velocity decreases and the optical density-of-states increases as the asymptote is approached. The SPP bend-back is also observed for wavelengths shorter than the asymptote ($\lambda_0 < 360$ nm, $E > 3.4$ eV) and links the nonradiative SPP on the right of the light line to the radiative one on the left.

The nonradiative SPP is located to the right of the light line (Fig. 2) and so cannot be directly excited by an incident beam. An additional structure is required to increase the in-plane momentum of the beam to match that of the SPP. A prism or a corrugated grating is commonly used to accomplish this task; alternatively, the SPP can be excited by end-fire coupling [1].

Detection Mechanisms

Figure 3a gives a schematic of a conventional Schottky photodiode on n-type Si (n-Si) [3],

adopted here to describe generic detection mechanisms that can also be used for the detection of SPPs. A rectifying Schottky contact is formed at the abrupt interface between the top metal and the semiconductor, and an Ohmic (non-rectifying) contact is formed at the interface between the bottom metal and the heavily doped body. The structure is reverse biased (for detection) by applying $V_b < 0$. The complementary structure, obtained by exchanging the dopants ($n \leftrightarrow p$, $n^+ \leftrightarrow p^+$), is also of interest. The Schottky diode is a convenient structure with which to detect SPPs because SPPs may propagate along its metal surfaces.

Figure 3a shows two mechanisms used for photodetection. The first consists of the creation of electron–hole pairs (EHPs) in the semiconductor due to absorption therein of incident radiation of energy $h\nu$ greater than the bandgap energy of the semiconductor E_g . This mechanism, sketched in Fig. 3a and labeled EHP, involves three steps: optical absorption and creation of EHPs, separation of EHPs and transport across the absorption region (with or without gain) under reverse bias, and collection of EHPs into the photocurrent at the device contacts.

The second mechanism consists of the internal photoemission (IPE) of hot carriers created in the metal due to absorption therein of incident

radiation of energy hv . This mechanism is sketched in Fig. 3a, labeled IPE, and is described in greater detail via the energy band diagram of Fig. 3b. IPE is a three-step process consisting of the photoexcitation of hot (energetic) carriers in the metal by optical absorption (p), the transport and scattering of hot carriers toward the Schottky contact (t), and the emission of hot carriers over the Schottky barrier into the semiconductor (e) where they are collected under a reverse bias as the photocurrent. This mechanism requires that hv be greater than the Schottky barrier energy Φ_B , and is useful when $\Phi_B < hv < E_g$; i.e., for detection at energies below the bandgap of the semiconductor (because the creation of EHPs in the semiconductor is more efficient and dominates over IPE when $hv > E_g$).

The internal quantum efficiency η_i (internal photoyield) is a useful measure to characterize and optimize the detection mechanism. It is defined as the number of carriers that contribute to the photocurrent I_p per absorbed photon per second:

$$\eta_i = \frac{I_p/q}{S_{abs}/hv} \quad (4)$$

where S_{abs} is the absorbed optical power and q is the elemental charge. $\eta_i \sim 1$ for detection via the creation of EHPs in high-quality (defect-free) direct bandgap semiconductors (assuming absorption in the semiconductor only). In the case of detection via IPE, assuming absorption in the metal only near the Schottky contact, η_i is given approximately by [4]:

$$\eta_i = \frac{1}{2} \left(1 - \sqrt{\frac{\Phi_B}{hv}} \right)^2 \quad (5)$$

Typical Schottky barrier heights are $\Phi_B = 0.34$, 0.8, 0.58, and 0.72 eV for Au/p-Si, Au/n-Si, Al/p-Si, and Al/n-Si, respectively [3]; for detection at, e.g., $\lambda_0 = 1,310$ nm η_i ranges from about 0.3 % to 9 %. Metal silicides can also be used, providing lower Schottky barriers.

The external quantum efficiency η_e (external photoyield) and the responsivity R describe how

well the detector performs when inserted into a system. η_e is defined similarly to η_i except that it depends on the incident optical power S_{inc} :

$$\eta_e = \frac{I_p/q}{S_{inc}/hv} \quad (6)$$

η_e and η_i are related by:

$$\eta_e = A\eta_i \quad (7)$$

where A is the optical absorptance, defined as:

$$A = \frac{S_{abs}}{S_{inc}} \quad (8)$$

The responsivity R is given by the ratio of the photocurrent to the incident optical power, and can be expressed in terms of η_e and η_i :

$$R = \frac{I_p}{S_{inc}} = \frac{\eta_e q}{hv} = \frac{A\eta_i q}{hv} \quad (9)$$

SPP detectors, or SPP-enhanced detectors, typically combine a metallic structure that supports SPPs with a detector structure such as that shown in Fig. 3.

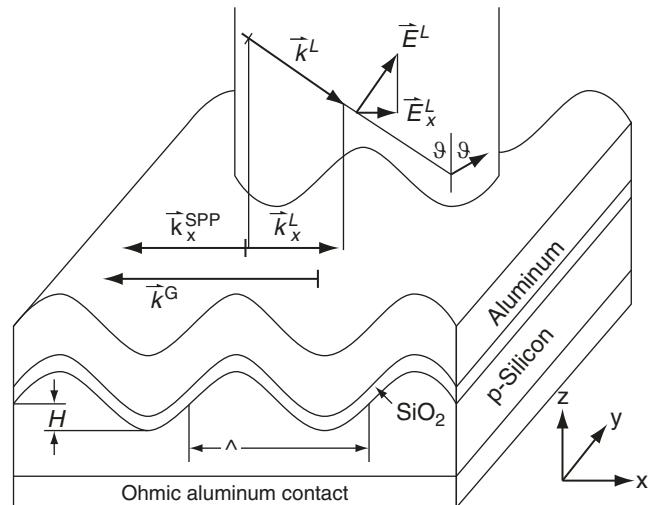
Grating-Coupled Detectors

An SPP detector structure of interest combines a Schottky detector with a grating designed to couple incoming optical waves to SPPs, as shown in Fig. 4 [5]. The structure consists of a top Al electrode on a thin SiO₂ tunneling barrier on p-Si, formed as a sinusoidal corrugated grating of height H , period A , and grating vector $\mathbf{k}^G = \mathbf{x}2\pi/A$ (\mathbf{x} is the x -directed unit vector). p-polarized light having a wavevector \mathbf{k}^L and an electric field \mathbf{E}^L is incident onto the grating at the angle ϑ . The incident light couples to the SPP propagating along the air/Al interface in the x direction when the following momentum conservation equation is satisfied:

$$k_x^L + m k^G = k_x^{spp} \quad (10)$$

Surface Plasmon-Polariton-Based Detectors, Fig. 4 SPP

Schottky detector integrated with a grating coupler to excite x -propagating SPPs along the top surface of the top Al contact (Adapted from Ref. [5]. © 1989 Optical Society of America)



where k_x^L is the x -directed component of \mathbf{k}^L , m is an integer, and k_x^{SPP} is the SPP wavenumber (following the notation of [5]). Equation 10 holds for shallow gratings – H is a small perturbation to the surface. The Al Schottky contact is thin enough (35 nm) for the SPP to tunnel through and leak into the high refractive index p-Si layer for detection via the creation of EHPs at the wavelength of operation ($\lambda_0 = 646$ nm).

Grating-coupled SPP detectors are sensitive to the angle of incidence, polarization, and wavelength of the incident light through Eq. 10. The polarization sensitivity, for instance, can be exploited in an arrangement of two detectors to determine the polarization angle of linearly polarized normally incident ($\vartheta = 0$) light [5].

Grating-coupled SPPs have also been used to increase the absorptance in the metal contact of Schottky detectors based on IPE (for sub-bandgap detection) leading to increased responsivity. For example, a $30\times$ increase in responsivity was observed for a Au/p-InP Schottky detector by exciting SPPs along the air/Au interface at $\lambda_0 = 1,150$ nm via an integrated corrugated grating, relative to the same structure without the grating [6].

Prism coupling, in the Otto arrangement [1] for instance, has also been used to excite SPPs on detector structures as an alternative to grating coupling. For example, a $10\text{--}20\times$ increase in IPE was observed for Au/n-GaAs Schottky

detectors by exciting SPPs along the Au/GaAs interface at $\lambda_0 = 1,150$ nm in an Otto arrangement implemented monolithically through the substrate [7].

Hole-Coupled Detectors

Optical transmission through a single sub-wavelength hole in a metal film can be significantly larger than predicted by Bethe's theory for a hole in a perfect conductor [8]. This larger (extraordinary) transmission is due to the excitation of SPPs on the metal surface near the hole that then propagate through the hole and couple to radiation on the other side; i.e., the transfer of energy through the hole is mediated by SPPs. Transmission at visible wavelengths is largest for a hole that is 150–300 nm in diameter in 200–300 nm thick Au or Ag films [8].

Structuring the metal film into a corrugation, e.g., around a sub-wavelength hole or slit leads to increased collection of incident light and to greater transmission at resonant wavelengths [8]. Applying this concept to detectors leads to devices that may offer improved performance [9–11].

Ishi et al. [9] reported a Schottky detector on n-Si where the metal forming the Schottky contact was structured into a circular corrugated grating surrounding a 300 nm diameter hole (and covered

by SiO₂). The contact metal was a Ag/Cr stack 200/10 nm thick. Detection was reported at 840 nm based on absorption in a 300 nm diameter n-Si mesa located directly below the hole, and the creation of EHPs therein. A prospective advantage of the structure is high-speed operation due to the small size of the detection volume while maintaining good responsivity due to the light collection ability of the grating/hole combination; thus the detector exhibits a favorable trade-off between these two parameters (compared to conventional detectors). It has also been predicted that the grating/hole (or slit) combination can lead to detectors having an improved signal-to-noise ratio [10, 11].

Arrays of holes in metal films also couple incident light to SPPs, in a manner analogous to a corrugated grating [8]. The momentum conservation equation for a two-dimensional (low perturbation) periodic structure on a metal surface is [8, 12]:

$$\mathbf{k}^{SPP} = \mathbf{k}_{\parallel} + l\mathbf{G}_x + m\mathbf{G}_y \quad (11)$$

where \mathbf{k}_{\parallel} is the in-plane wavevector of the incident light, \mathbf{G}_x and \mathbf{G}_y are reciprocal lattice vectors of the periodic structure, and l, m are integers. This equation models approximately the coupling of light to SPPs on a two-dimensional hole array, although the holes are actually neglected along with their associated effects such as scattering and transmission. For example, in the case of a hexagonal or triangular lattice of holes of lattice constant a , illuminated at normal incidence ($\mathbf{k}_{\parallel} = 0$) and neglecting all losses, the above yields the wavelengths λ_c for which strong coupling to SPPs on the array occurs:

$$\lambda_c = a \left[\frac{4}{3} (l^2 + lm + m^2) \right]^{\frac{1}{2}} \left(\frac{\varepsilon_R \varepsilon_{r,d}}{\varepsilon_R - \varepsilon_{r,d}} \right)^{\frac{1}{2}} \quad (12)$$

In the above, $\varepsilon_{r,d}$ corresponds to the dielectric bounding the array on the illuminating side. If the structure is symmetric (same dielectric on both sides of the array), then these wavelengths also correspond to those at which transmission peaks occur. If the structure is asymmetric

(different dielectrics bounding the array) then two sets of transmission peaks are observed.

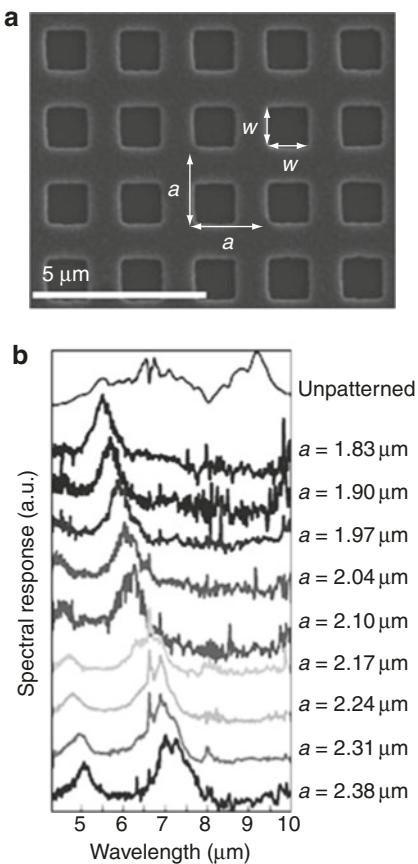
Hole arrays can be integrated with broadband detectors to introduce wavelength and polarization selectivity, and potentially, to enhance the responsivity [12, 13]. This is achieved by structuring the top or bottom contact of a detector into an array of sub-wavelength holes and then illuminating at normal incidence. The photocurrent is then largest at wavelengths for which extraordinary transmission occurs, which depend on the lattice constants of the array, e.g., via Eq. 12 for the case of a hexagonal array, and so are easily tunable. Polarization selectivity can be introduced by using elliptical or rectangular holes.

The responsivity can be enhanced for a thin detection layer by enhancing the absorptance (A) via mediation by SPPs. This is achieved by placing the detection layer close to the hole array such that SPPs propagating along the array overlap with the layer and are absorbed therein; thus the enhancement arises because SPPs propagate along the absorption region rather than through it perpendicularly.

Wavelength selectivity and responsivity enhancement are particularly attractive for mid-infrared detector arrays, which are typically broadband and have a low responsivity due to the low absorption of the detector materials used in this wavelength range. Applications such as multi-color (multi-spectral) night vision or medical imaging would benefit from inexpensive spectral filtering integrated at the pixel level. Figure 5a shows a square array (lattice constant a) of square holes (dimensions w) patterned onto the top contact (150-nm thick Ag) of an InAs mid-infrared quantum dot detector [13]. Figure 5b shows the measured spectral response of such detectors as a function of a , compared to an unpatterned control detector (upper trace). The peak wavelength response is observed to range from $\lambda_0 = 5.5\text{--}7.2 \mu\text{m}$ as the lattice constant ranges from $a = 1.83\text{--}2.38 \mu\text{m}$.

Detectors Incorporating Nanoparticles

Small metal particles exhibit resonant responses under optical excitation, characteristic of particle



Surface Plasmon-Polariton-Based Detectors,
Fig. 5 (a) Image of a square lattice of square holes; the lattice constant is a and the hole size is w . (b) Spectral response of quantum dot detectors each bearing a hole array of the indicated lattice constant; $w = 0.6a$ for all detectors (Adapted from Ref. [13]. © 2009 American Institute of Physics)

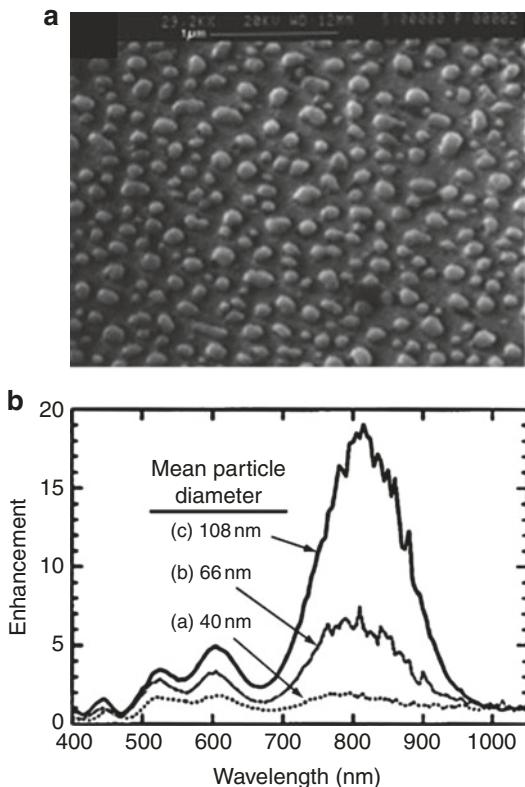
shape, size, and composition, the dielectric environment in which they find themselves, and the wavelength of illumination [14]. Resonances occur when the electron charge density oscillates coherently with the illuminating optical (electrical) field. The fundamental resonant mode of, e.g., a spherical metal nanoparticle (small compared to the illuminating wavelength) is dipolar with densities of opposite charge forming at opposite spherical caps of the particle along the polarization of the illuminating electric field. The wavelength of the dipolar resonance red-shifts as the radius of the particle increases or as the index of the background increases. If the

particle is large enough then higher-order resonant modes can exist (e.g., quadrupolar). On resonance, the electric field near the particle (in the dielectric) is strongly enhanced compared to the illuminating field. Resonant features appear in the measured extinction and scattering spectra of metal nanoparticles, in correlation with the plasmon resonances that are excited thereon.

Metal nanoparticles can be integrated with a detector most readily via deposition onto the surface through which light penetrates into the detector [15, 16]. As light propagates through, resonances can be excited, and strong scattering can occur, leading to improved detector performance.

Figure 6a shows a scanning electron microscope image of Ag-island nanoparticles on a 30-nm thick LiF layer on a Si-on-insulator (SOI) detector [15]. These approximately hemispherical particles are 108 nm across, on average. They were formed by first depositing a thin Ag film, then annealing the film to induce islandization. This process has the advantage of being simple but it generates particles that are non-uniform in shape and size, albeit, with a controllable average size. Detectors were formed in the thin Si slab (165 nm thick) above the insulator but beneath the LiF (and the nanoparticles) as pn junction detectors. Figure 6b shows measured spectral responses of detectors coated with Ag islands similar to those of Fig. 6a as a function of average island size. The enhancement is defined as the ratio of the photocurrent from a detector with Ag islands to the photocurrent from a detector without islands. A strongly enhanced photocurrent ($\sim 20\times$) is noted at $\lambda_0 = 800$ nm in the case of islands that are 108 nm in average size. The enhancement is attributed to the large scattering cross-section of the islands, particularly to mediation by the islands whereby dipole resonances excited thereon radiate into guided modes of the thin underlying Si slab. These modes propagate longitudinally and thus are absorbed more effectively by the detector compared to perpendicularly incident light.

Chemically synthesized (colloidal) nanoparticles can also be deposited and attached to detector surfaces. Spectral responses of



Surface Plasmon-Polariton-Based Detectors,
Fig. 6 (a) Scanning electron microscope image of Ag-island nanoparticles on LiF-coated Si-on-insulator (SOI) detector. The particles are 108 nm across, on average. (b) Spectral response of pn junction SOI photodetectors coated with Ag islands similar to those of (a) as a function of average island size. The enhancement is defined as the ratio of the photocurrent from a detector with Ag islands to the photocurrent from a detector without the islands (Adapted from Ref. [15]. © 1998 American Institute of Physics)

detectors coated with 50, 80, and 100 nm diameter spherical Au nanoparticles showed strong photocurrent enhancement at wavelengths where peaks were measured in the extinction spectra of the particles [16]. In contrast to the devices of [15], the enhanced photocurrent was attributed to enhanced absorption in the Si regions in contact with resonating nanoparticles, around which the resonating electric fields are significantly enhanced compared to the illuminating fields.

The effects induced by metal nanoparticles, i.e., increased scattering into detectors and increased detector absorption on resonance, are

useful to increase the absorptance of detectors, particularly those based on a thin absorption layer, or at wavelengths where the absorption is low (e.g., near the bandgap of an indirect semiconductor). These effects can potentially improve the efficiency of, e.g., thin-film Si solar cells [17].

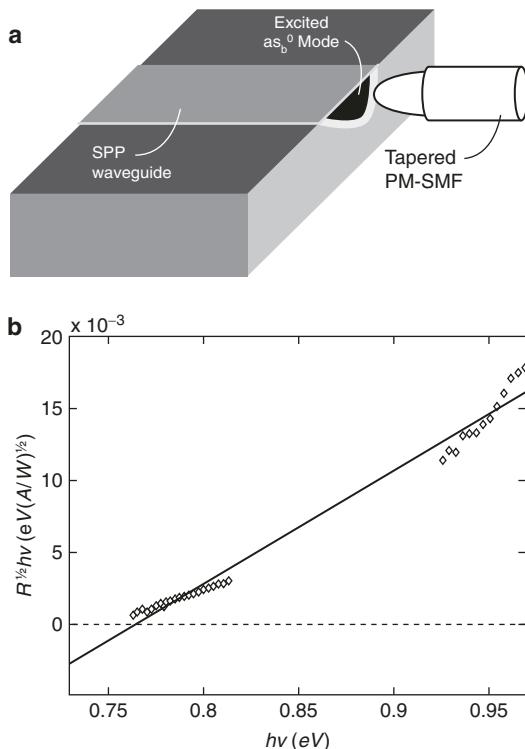
Waveguide Detectors

The single-interface (Fig. 1) is one example of an SPP waveguide; other popular 1-D geometries include the thin metal slab bounded by dielectrics and the thin dielectric slab bounded by metals [2]. The latter are often termed the IMI and MIM, respectively (I – insulator, M – metal). The IMI and MIM support super-modes formed from the coupling of single-interface SPPs through the thin intervening layer. The structures can be designed such that they support super-modes that are low loss but weakly confined (symmetric IMI) or strongly confined but high loss (MIM), and thus that are at opposite ends of the confinement-attenuation trade-off [2].

SPP waveguides can be formed into SPP detectors and conveniently integrated with other plasmonic or photonic waveguide structures. SPP waveguide detectors based on absorption in, e.g., organics [18], semiconductors [19], or metals [20] have been reported.

The detector described in [18] consists of a thick Ag film on which SPPs are excited by illuminating a slit in the film. The SPPs then propagate along the Ag film to the detector section, which consists of a pn junction fabricated from organic polymers directly on the Ag film. The pn junction is covered by another Ag film, thereby creating an MIM-type detector. SPPs incident from the Ag film onto this MIM detector excite SPPs therein, which are absorbed by the pn junction creating EHPs that are collected as the photocurrent. Current maps correlating the photocurrent with the excitation of SPPs at $\lambda_0 = 632.8$ nm along the Ag film are reported.

The structure described in [19] consists of an MIM (Au/HSQ/Au) waveguide with slits, fabricated on a GaAs wafer. A slit in the top Au layer is used to couple incident light to SPPs propagating



Surface Plasmon-Polariton-Based Detectors,
Fig. 7 (a) Sketch of an SPP waveguide detector consisting of a metal stripe forming a Schottky contact on Si, excited (in the as_b^0 mode) via butt-coupling to a tapered polarization-maintaining single mode fiber (PM-SMF). (b) Spectral response of a Au on n-Si detector (as in (a)), plotted in Fowler form (Adapted from n. © 2010 Optical Society of America)

in the MIM. A slit in the bottom Au layer, placed a short distance away from the input, couples the SPP into radiation directed into the GaAs wafer, which is absorbed therein creating EHPs that are collected as the photocurrent. Spectral responses and photocurrent maps demonstrating the operation of the detector are reported.

Figure 7a gives a sketch of a SPP waveguide detector consisting of a thin narrow metal stripe on semiconductor (with air on top) [20]. The thin metal stripe is similar to the asymmetric IMI, but it supports SPPs with additional confinement along the lateral dimension (stripe width). Detectors have been fabricated as Au or Al stripes on n-Si forming Schottky contacts thereon [20]. The as_b^0 mode, localized to the bottom metal-Si interface, is excited via butt-coupling to a tapered

polarization-maintaining single mode fiber (PM-SMF). This mode propagates along the stripe with strong absorption, creating hot carriers therein, some of which may cross the Schottky barrier and be collected as the photocurrent; thus detection occurs via IPE. Figure 7b gives the spectral response of a Au on n-Si detector, plotted in Fowler form. The intercept with the abscissa yields the cutoff photon energy (~ 0.765 eV), corresponding in this case to a cutoff wavelength of $\lambda_0 \sim 1,620$ nm. Responsivities up to 1 mA/W were reported with this detector scheme.

Concluding Remarks

SPP detectors, or SPP-enhanced detectors, typically combine a metallic structure that supports SPPs, such as a planar or grating structure, metallic nanoparticles, or holes in metal films, with a semiconductor detector structure such as a Schottky or a pn junction. Properties inherent to SPPs are exploited to convey additional characteristics to a detector such as polarization, angular or spectral selectivity, or to enhance the absorptance (A) of the detector. The involvement of SPPs has led to detectors with improved performance and greater functionality.

References

1. Maier, M.A.: *Plasmonics: Fundamentals and Applications*. Springer, New York (2007)
2. Berini, P.: Figures of merit for surface plasmon waveguides. *Opt. Express* **14**, 13030–13042 (2006)
3. Sze, S.M.: *Physics of Semiconductor Devices*. Wiley, New York (1981)
4. Scales, C., Berini, P.: Thin-film Schottky barrier photodetector models. *IEEE J. Quantum Electron.* **46**, 633–643 (2010)
5. Jestl, M., Maran, I., Kock, A., Beinstingl, W., Gornik, E.: Polarization-sensitive surface plasmon Schottky detectors. *Opt. Lett.* **14**, 719–721 (1989)
6. Brueck, S.R.J., Diadiuk, V., Jones, T., Lenth, W.: Enhanced quantum efficiency internal photoemission detectors by grating coupling to surface-plasma waves. *Appl. Phys. Lett.* **46**, 915–917 (1985)
7. Daboo, C., Baird, M.J., Hughes, H.P., Apsley, N., Emeny, M.T.: Improved surface plasmon enhanced photodetection at an Au-GaAs Schottky junction

- using a novel molecular beam epitaxy grown Otto coupling structure. *Thin Solid Films* **201**, 9–27 (1991)
8. Genet, G., Ebbesen, T.W.: Light in tiny holes. *Nature* **445**, 39–46 (2007)
 9. Ishii, T., Fujikata, J., Makita, K., Baba, T., Ohashi, K.: Si Nano-photodiode with a surface plasmon antenna. *Jpn. J. Appl. Phys.* **44**, L364–L366 (2005)
 10. Yu, Z., Veronis, G., Fan, S., Brongersma, M.L.: Design of midinfrared photodetectors enhanced by surface plasmons on grating structures. *Appl. Phys. Lett.* **89**, 151116 (2006)
 11. Bhat, R.D.R., Panoiu, N.C., Brueck, S.J.R., Osgood, R.M.: Enhancing the signal-to-noise ratio of an infrared photodetector with a circular metal grating. *Opt. Express* **16**, 4588–4596 (2008)
 12. Chang, C.Y., Chang, H.Y., Chen, C.Y., Tsai, M.W., Chang, Y.T., Lee, S.C., Tang, S.F.: Wavelength selective quantum dot infrared photodetector with periodic metal hole arrays. *Appl. Phys. Lett.* **91**, 163107 (2007)
 13. Rosenburg, J., Shenoi, R.V., Vandervelde, T.E., Krishna, S., Painter, O.: A multispectral and polarization-selective surface-plasmon resonant midinfrared detector. *Appl. Phys. Lett.* **95**, 161101 (2009)
 14. Kelly, K.L., Coronado, E., Zhao, L.L., Schatz, G.C.: The optical properties of metal nanoparticles: the influence of size, shape, and dielectric environment. *J. Phys. Chem. B* **107**, 668–677 (2003)
 15. Stuart, H.R., Hall, D.G.: Island size effects in nanoparticle-enhanced photodetectors. *Appl. Phys. Lett.* **73**, 3815–3817 (1998)
 16. Schaadt, D.M., Feng, B., Yu, E.T.: Enhanced semiconductor optical absorption via surface plasmon excitation in metal nanoparticles. *Appl. Phys. Lett.* **86**, 063106 (2005)
 17. Atwater, H.A., Polman, A.: Plasmonics for improved photovoltaic devices. *Nat. Mater.* **9**, 205–213 (2010)
 18. Ditlbacher, H., Aussenegg, F.R., Krenn, J.R., Lamprecht, B., Jakopic, G., Leising, G.: Organic diodes as monolithically integrated surface plasmon polariton detectors. *Appl. Phys. Lett.* **89**, 161101 (2006)
 19. Neutens, P., Van Dorpe, P., De Vlaminck, I., Lagae, L., Borghs, G.: Electrical detection of confined gap plasmons in metal-insulator-metal waveguides. *Nat. Photonics* **3**, 283–286 (2009)
 20. Akbari, A., Tait, R.N., Berini, P.: Surface plasmon waveguide Schottky detector. *Opt. Express* **18**, 8505–8514 (2010)

Surface Properties

- Nanostructures for Surface Functionalization and Surface Properties

Surface Science

- Influence of Defects in Photocatalysis

Surface Tension and Chemical Potential at Nanoscale

- Surface Energy and Chemical Potential at Nanoscale

Surface Tension Effects of Nanostructures

Ya-Pu Zhao and Feng-Chao Wang
State Key Laboratory of Nonlinear Mechanics (LNM), Institute of Mechanics, Chinese Academy of Sciences, Beijing, China

Synonyms

Interface excess free energy; Surface energy density

Definition

The surface tension is the reversible work per unit area needed to elastically stretch/compress a preexisting surface. In other words, it is a property of the surface that characterizes the resistance to the external force. Surface tension has the dimension of force per unit length or of energy per unit area. For nanostructures with a large surface to volume ratio, the surface tension effects dominate the size-dependent mechanical properties.

Overview

Nanostructures have a sizable surface to volume ratio as compared to bulk materials, which leads

its mechanical properties to be quite different from those of bulk materials [1]. The size-dependent mechanical properties of nanostructures are generally attributed to the surface effects, in which surface tension is one of the most predominant factors.

In the framework of the thermodynamics, the surface (interface) between two phases is first modeled as a bidimensional geometrical boundary of zero thickness, that is, the mathematical surface, as shown in Fig. 1a. The physical quantities between the two phases are discontinuous, and the surface (interfacial) tension is a jump in stress. This idealization was then extended by Gibbs [2], who proposed that the physical quantities should undergo a smooth transition at the surface (interface) while the surface is still modeled as an infinitesimal thin boundary layer, as shown in Fig. 1b. In order to preserve the total physical properties of the system, the excess physical properties have to be assigned to the geometrical surface. The surface tension, which is defined based on the interface excess free energy, can be written as

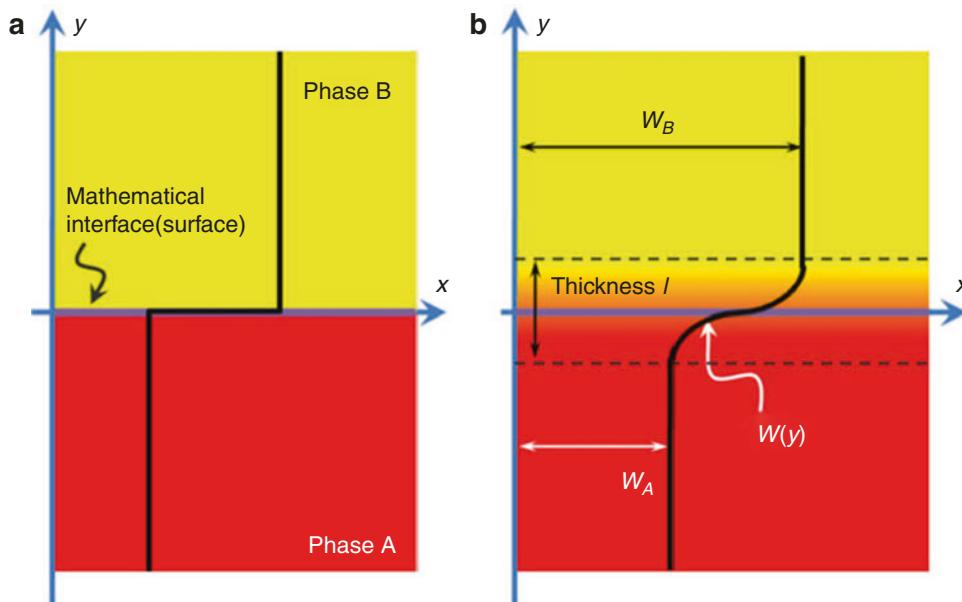
$$\gamma = \int_0^\infty [w(y) - W_A] dy + \int_{-\infty}^0 [w(y) - W_B] dy \quad (1)$$

where w is the free energy distribution of the actual surface, w_A and w_B are free energy in the two phases of A and B .

In some other theoretical models, the surface is treated as an extended interfacial region with a nonzero thickness. According to Cahn–Hilliard theory [3], the surface tension can be derived as

$$\gamma = N_V \int_{-\infty}^\infty [w_0(c) + k(dc/dy)^2 - c\mu_B - (1-c)\mu_A] dy, \quad (2)$$

in which N_V is the number of atoms per unit volume, c is one of the intensive scalar properties, such as composition or density, $w_0(c)$ is the free energy per atom of a solution of uniform composition c , k reflects the crystal symmetry, μ_A and μ_B are the chemical potentials per atom in the



Surface Tension Effects of Nanostructures, Fig. 1 An illustration for the surface (interface) models. (a) The mathematical surface of zero thickness, the physical quantities are discontinuous. (b) Gibbs's surface with an

infinitesimal thickness, the physical quantities are continuous. Cahn–Hilliard model for surface with a finite thickness is also illustrated

A or *B* phase. The surface (interfacial) thickness can be obtained by

$$l = 2\Delta c_e \sqrt{\frac{k}{\Delta w_{\max}}}, \quad (3)$$

where $2\Delta c_e = c_B - c_A$ is the difference of the uniform composition in the two phases, Δw_{\max} is the maximum of the free energy referred to a standard state of an equilibrium mixture. Cahn–Hilliard model gives a finite thickness through Eq. 3, which has been used in many applications [4].

From the standpoint of molecular theory, the surface tension effects arise due to the difference of the atomic interactions in the bulk and on the surface. Atoms are energetically favorable to be surrounded by others. At the surface, the atoms are only partially surrounded by others and the number of the adjacent atoms is smaller than in the bulk. Thus the atoms at the surface are energetically unfavorable. If an atom moves from the bulk to the surface, work has to be done. With this view, the surface tension can be interpreted as the energy required to bring atoms from the bulk to the surface [5]. Therefore the term “surface energy density” is often used to when the surface tension is referred to. For the surface of solid, Gibbs pointed out that surface tension and surface energy density are not identical. The surface tension is the reversible work per unit area needed to elastically stretch/compress a preexisting surface. The surface energy density is the reversible work per unit area needed to create a new surface. The surface tension can be positive or negative, while the surface energy density is usually positive. The surface tension for liquid surface is a property that characterizes its resistance to the external force, which is identical to the surface energy density.

Basic Methodology

Since the surface to volume ratio increases as the dimension scale decreases, surface tension effects of nanostructures can be overwhelming. In the absence of external loading, the surface tension

effects would induce a residual stress field in bulk materials. The relations between surface stress and surface tension for small deformations can be described by the Shuttleworth-Herring equation [6, 7],

$$\sigma_{ij} = \gamma \delta_{ij} + \frac{\partial \gamma}{\partial \varepsilon_{ij}}, \quad (4)$$

where γ is the surface tension, δ_{ij} is the Kronecker delta, σ_{ij} and ε_{ij} are the surface stress tensor and the surface strain tensor, respectively. The Shuttleworth-Herring equation interprets that the difference between the surface stress and surface tension is equal to the variation of surface tension with respect to the elastic strain of the surface.

With the development of computational materials science, molecular dynamics (MD) simulations are wildly performed to investigate the surface tension effects on the mechanical properties of nanostructures. Especially for the surface elastic constant, atomistic simulation is almost the only way to get them up to now. According to the Gibbs's definition, the surface tension of a solid is given by $\gamma = (E_S - nE_B)/A_0$, where A_0 is the total area of the surface considered, E_S is the total energy of a n -layer slab and E_B is the bulk energy per layer of an infinite solid. In cases where the surfaces of the slab are polar, the electrostatic energy of the slab contains an energy contribution, E_{pol} , proportional to the substrate thickness and the surface energy needs corrections. Thus there is an alternative way to calculate the surface tension $\gamma = (E_S - nE_B - E_{\text{pol}})/A_0$, which does not rely on an exact knowledge of the lattice or polar energy. MD simulations have identified that the surface tension induced surface relaxation is proved to be a dominate factor of the size-dependent mechanical properties. When the surface stress is negative, the surface relaxation is inward; otherwise, the relaxation is outward.

Surface stress has been used as an effective molecular recognition mechanism. Surface stresses due to DNA hybridization and receptor-ligand binding induce the deflection of a cantilever sensor [8]. The curvature of bending beam

under a surface stress is governed by Stoney's formula [9]. Stoney's formula serves as a cornerstone for curvature-based analysis and a technique for the measurement of surface stress, which is given as follows as a general form for a film/substrate system,

$$\sigma = \frac{Et_s^2 f}{6(1 - v)} \quad (5)$$

in which E is the effective Young's modulus, v is Poisson's ratio of the sensor material, t_s is the substrate thickness, $f = 3\Delta z/2 L^2$ is the sensor curvature, L is the length and Δz is the deflection. The applicability of the above Stoney's formula relies on several assumptions, which are well summarized as the following six: (1) both the film and substrate thicknesses are small compared to the lateral dimensions; (2) the film thickness is much less than the substrate thickness; (3) the substrate material is homogeneous, isotropic, and linearly elastic, and the film material is isotropic; (4) edge effect near the periphery of the substrate are inconsequential and all physical quantities are invariant under change in position parallel to the interface; (5) all stress components in the thickness direction vanish throughout the material; and (6) the strains and rotations are infinitesimally small. However, the one or several of above six assumptions can be easily violated in reality, which is to say that the Stoney's formula needs to be revised to fit in the real applications.

To summarize for the solid cases, the behaviors of nanostructures can be affected significantly by either of the two distinct parameters, surface tension and surface stress. The relation between the two parameters can be obtained by the Shuttleworth-Herring equation. For the surface stress induced deflection of cantilever sensors, the curvature is by Stoney's formula. MD simulation is helpful to understand the surface effects of nanostructures and partial results are comparable to the experiments.

For liquid droplets in contact with the surface of a nanostructure, surface tension is responsible for the shape of liquid droplets in the equilibrium state, as well as the dynamics response in wetting

and dewetting. There are several characteristic time which are related to the surface tension effects, listed in Table 1. Dimensionless number related to the surface tension effects are listed in Table 2. Surface tension is dependent on temperature T , concentration of surfactants c , and the electric field V . The gradient of surface tension caused by these factors can be described by [10]

$$d\gamma = \frac{\partial\gamma}{\partial T}dT + \frac{\partial\gamma}{\partial c}dc + \frac{\partial\gamma}{\partial V}dV. \quad (6)$$

To the first order, the dependency of the surface tension on temperature is given by Guggenheim-Katayama formula with power index $n = 1$, $\gamma = \gamma_0(1 - T/T_c)$. Here γ_0 is a constant for each liquid and T_c is the critical temperature. For the dependency of the surface tension on concentration c , the surface tension can be expressed as a linear function of the concentration, $\gamma = \gamma_0[1 + \beta(c - c_0)]$. The solid–liquid surface tension can be changed by applying a voltage V , $\gamma_{SL} = \gamma_{SL}^0 - \frac{\epsilon_0 \epsilon_D}{2d} V^2$, where γ_{SL}^0 is the solid–liquid surface tension in the absence of the applied voltage, ϵ_0 is the permittivity of vacuum, ϵ_D is the relative permittivity of the dielectric layer

Surface Tension Effects of Nanostructures, Table 1 Characteristic time related to the surface tension effects

Name	Expression	Meaning
Capillary characteristic time	$t_c = \sqrt{m/\gamma_{LV}}$	Characteristic time derived from the droplet mass and the liquid–vapor surface tension
Lord Rayleigh's period	$t_p = \frac{\pi}{4} \sqrt{\rho d_0^3 / \gamma_{LV}}$	The period of a free droplet in free oscillation
Lord Rayleigh's characteristic time	$t_R \sim \sqrt{\rho l^3 / \gamma_{LV}}$	Characteristic time of droplet dynamics
Viscous characteristic time	$t_{vis} \sim \eta l / \gamma_{LV}$	Characteristic time related to the liquid viscosity

m mass of the droplet, d_0 diameter, ρ density, η viscosity, l characteristic length

Surface Tension Effects of Nanostructures, Table 2 Dimensionless number related to the surface tension effects

Name	Expression	Meaning
Adhesion number	$N_a = \cos \theta_a - \cos \theta_r$	θ_a and θ_r are the advancing and receding angle
Bond number	$Bo = \rho g^2 / \gamma_{LV}$	The capillary length $l_c = \sqrt{\gamma_{LV}/(\rho g)}$ can be determined
Capillary number	$Ca = \eta v / \gamma_{LV}$	$Ca = On \cdot \sqrt{We}$; the capillary velocity $v = \gamma_{LV}/\eta$ can be obtained
Deborah number	$De = t/t_R = t / \sqrt{\rho l^3 / \gamma_{LV}}$	t is the relaxation time, Lord Rayleigh characteristic time can be derived
Elasto-capillary number	$Ec = t\gamma_{LV} / (\eta l)$	t is the relaxation time, viscous characteristic time can be obtained
Electrowetting number	$\eta_e = \epsilon_0 \epsilon_D V^2 / (2d\gamma_{LV})$	The ratio of the electrostatic energy to the surface tension
Laplace number	$La = (1/On)^2$	See Ohnersoge number
Marangoni number	$Ma = \frac{d\gamma_{LV}}{dt} \frac{L \Delta T}{\eta \alpha}$	Thermal surface tension force divided by viscous force
Ohnersorge number	$On = \eta / \sqrt{\rho l \gamma_{LV}}$	$\frac{\text{Viscousforce}}{\sqrt{\text{Inertiaforce} \times \text{Surfacetension}}}$
Weber number	$We = \rho v^2 l / \gamma_{LV}$	The inertia force compared to its surface tension

g gravitational acceleration, α thermal diffusivity, ΔT temperature difference

with a thickness d separating the bottom electrode from the liquid.

The wetting properties of a solid surface can be described by the introduction of the contact angle, which is defined as the angle at which the liquid–vapor interface meets the solid surface. The contact angle is affected by various factors, including the surface tension, the line tension, the applied voltage as well as the molecular interactions between the liquid and solid surfaces. It is proposed that the dependence of the contact angle on these factors can be represented by the generalized Young's equation, which has a form of

$$\cos \theta = \cos \theta_0 - \frac{\tau}{\gamma_{LV} R} + \frac{\epsilon_0 \epsilon_D}{2d\gamma_{LV}} V^2 + \frac{A}{12\pi h^2 \gamma_{LV}}, \quad (7)$$

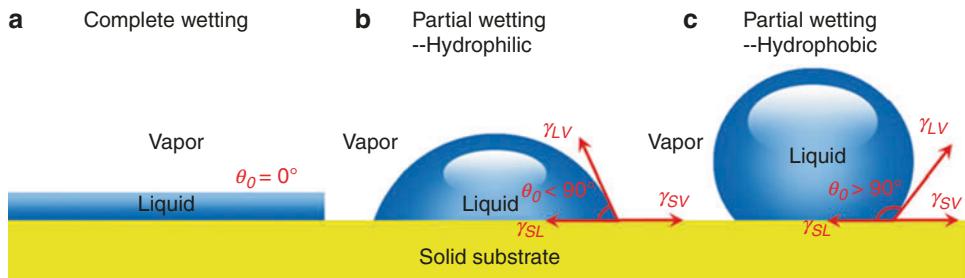
where

$$\cos \theta_0 = \frac{\gamma_{SV} - \gamma_{SL}}{\gamma_{LV}} \quad (8)$$

is the classical Young's equation, θ_0 is the equilibrium contact angle (also the Young contact angle); The subscripts S , L , and V denote solid, liquid, and vapor, respectively. The second term on the right-hand side of Eq. 7 is related to the line

tension. τ is the line tension and R is the radius of the contact area. For a more generalized case, R can be replaced by $1/\kappa$, in which κ is the geodesic curvature of the triple contact line. $\eta_e = \frac{\epsilon_0 \epsilon_D}{2d_{LVd}} V^2$ is defined as the dimensionless electrowetting number. The last term in Eq. 7 measures the strength of the effective interaction energy compared to surface tension. The interaction energy is related to the disjoining pressure $\Pi(h)$, which can be obtained by $W(h) = \int_h^\infty \Pi(h') dh'$, where h is the film thickness and A is the Hamaker constant [11, 12]. As discussed in the following paragraphs, each term would be illustrated in detail.

If only the classical Young's equation is referred to, the wetting properties of the surface can be obtained directly if the three tensions are known. The spreading parameter S determines the type of spreading, which is defined as $S = \gamma_{SV} - (\gamma_{SL} + \gamma_{LV})$. If $S > 0$, the liquid spreads on the solid surface and forms a macroscopic liquid layer covers the whole solid surface; it is the case for complete wetting. If $S < 0$, the contact angle $0^\circ < \theta_0 < 180^\circ$, which means the liquid forms a droplet with a finite contact angle; it is the case for speak of partial wetting. The wettability of the solid surface could be distinguished by the contact angle θ_0 , as shown in Fig. 2. If



Surface Tension Effects of Nanostructures, Fig. 2 Wetting properties of the surface

$0^\circ \leq \theta_0 < 90^\circ$, the solid substrate is hydrophilic. If $90^\circ < \theta_0 \leq 180^\circ$, the solid substrate is hydrophobic. Especially, if $150^\circ < \theta_0 \leq 180^\circ$, the solid substrate is superhydrophobic.

The physics behind wettability is that, the solid surfaces have been divided into high energy and low-energy types [11, 12]. The relative energy of a solid has to do with the bulk nature of the solid itself. (1) High-energy surfaces such as metals, glasses, and ceramics are bound by the strong chemical bonds, for example, covalent, ionic, or metallic, for which the chemical binding energy E_{binding} is of the order of 1 eV. The solid–liquid interface tension is given by $\gamma_{SV} \approx E_{\text{binding}}/a^2 \sim 0.5 - 5 \text{ N/m}$, in which a^2 is the effective area per molecule. Most high-energy surfaces are hydrophilic, some can permit complete wetting. (2) For low-energy surfaces, such as weak molecular crystals (bound by van der Waals forces or in some special cases, by hydrogen bonds), the chemical binding energy is of the order of $k_B T$. In this category, the surface tension is $\gamma_{SV} \approx k_B T/a^2 \sim 0.01 - 0.05 \text{ N/m}$. Depending on the type of liquid chosen, low-energy surfaces can be either hydrophobic or hydrophilic.

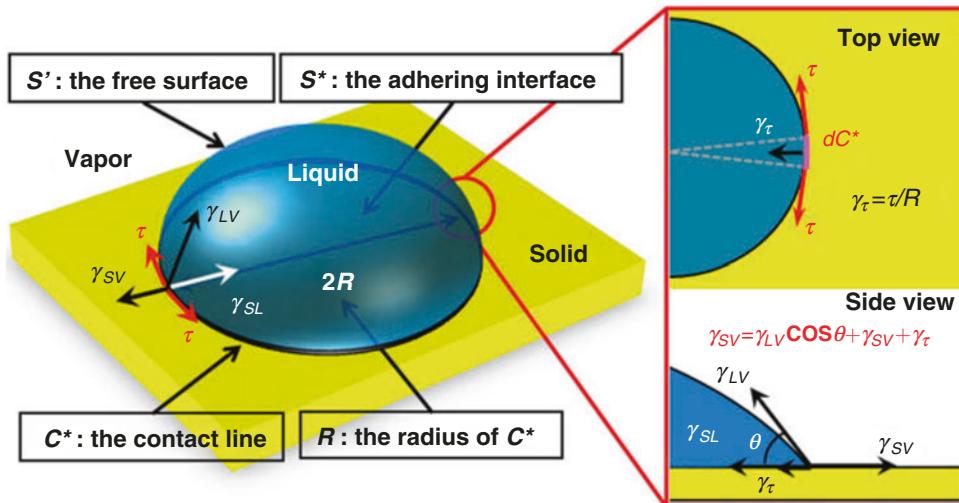
When the liquid droplet comes to the nanoscale, the classical Young's equation seems to be not applicable, since it has been derived for a triple line without consideration of the interactions near the triple contact line. The molecules close to the triple line experience a different set of interactions than at the interface [10]. To take into account this effect, the “line tension” term has been introduced in the generalized Young's equation. A sketch of line tension is shown in Fig. 3.

Line tension was first introduced by Gibbs [2]. The line tension τ was introduced as an analogue of the surface tension:

$$F = \gamma_{LV} \int_{S'} dS' + (\gamma_{LV} - W) \int_{S'} dS^* + \tau \int_{C^*} dC^*, \quad (9)$$

where F , S' , S^* , and C^* are the free energy, the free surface, the adhering interface, and the contact line of the droplet. $w = \gamma_{LV} + \gamma_{SL} - \gamma_{SV}$ is the adhesion potential. The line tension, depending on the radius R of the contact line, should be connected to the classical Young's equation to include the interactions near the triple contact line. In the three-phase equilibrium systems when R decreases, the contact angle θ will increase at $\tau > 0$ and will decrease at $\tau < 0$. However, in accordance to the mechanical equilibrium stability conditions, while the surface tension can only be positive, the line tension can have either positive or negative values. The characteristic length of this problem $l = |\tau|/\gamma_{LV}$ ($|\tau| < 10^{-9} \text{ N}$, $\gamma_{LV} \sim 0.1 \text{ N/m}$ for water) ranges from 10^{-8} to 10^{-6} m , which means small droplet (with typical dimension of $|l|$) should appreciate the line tension effect. The line tension can be indirectly measured through the contact angle, $\tau \approx 4\delta\sqrt{\gamma_{SV}\gamma_{LV}} \cot\theta$, in which δ denotes the average distance between liquid and solid molecules. Thus the line tension is negative for an obtuse contact angle, while it is positive for an acute contact angle.

In the presence of a charged interface, which can be achieved by applying a direct or



Surface Tension Effects of Nanostructures, Fig. 3 Illustration of the line tension τ

alternating-current electric field, the wetting properties of solid surface will be modified. The physics describing the electric forces on interfaces of conducting liquids and on triple contact lines is called “electrowetting” [10]. In the year of 1875, Gabriel Lippmann observed the capillary depression of mercury in contact with an electrolyte solution could be varied by applying a voltage between the mercury and electrolyte. This phenomenon is called electrocapillarity, which is the basis of modern electrowetting. Then, the idea was developed to isolate the liquid droplet from the substrate using a dielectric layer in order to avoid electrolysis. This concept has subsequently become known as electrowetting on dielectric (EWOD) and involves applying a voltage to modify the wetting behavior of a liquid in contact with a hydrophobic, insulated electrode. When an electric field was applied to the system (as shown in Fig. 4), electric charges gather at the interface between the conductive electrodes and the dielectric material; the surface becomes increasingly hydrophilic (wettable), the contact angle is reduced and the contact line moves. The change in contact angle over the buried electrodes can be evaluated by the Lippmann–Young equation. The parabolic variation of $\cos \theta$ in the Lippmann–Young equation is $\cos \theta = \cos \theta_0 + \epsilon_0 \epsilon_D V^2 / (2d\gamma_{LV})$

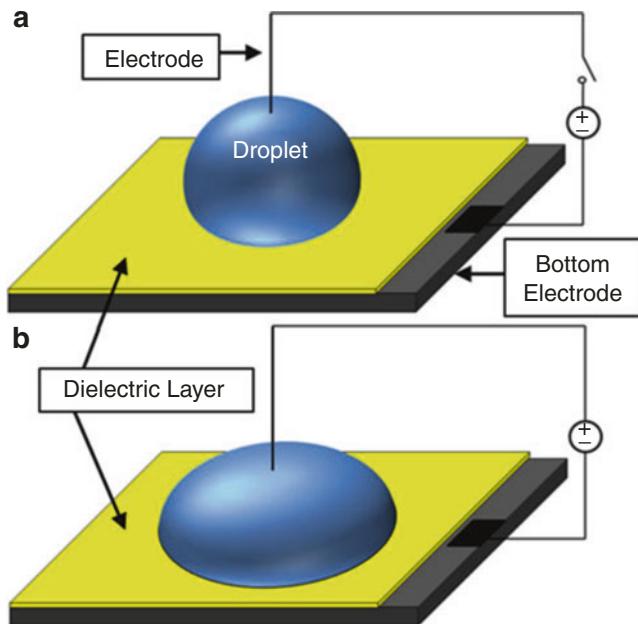
only applicable when the applied voltage lies below a threshold value. If the voltage is increased above this threshold, then contact angle saturation starts to occur and $\cos \theta$ eventually becomes independent of the applied voltage.

The physics of why a liquid film wets or dewets is found in the derivative of the effective interfacial potential $W = \gamma_{LV} + \gamma_{SL} - \gamma_{SV}$ with respect to film thickness h , called the disjoining pressure $\Pi(h)$ [11, 12]. $\Pi(h) = dW(h)/dh$, where the surface area remains constant in the derivative. The effective interface potential $W(h)$, which arises from the interaction energies of molecules in a film being different from that in the bulk, is the excess free energy per unit area of the film. If the interactions between the molecules in the film and the solid substrate are more attractive than the interactions between molecules in the bulk liquid, $W(h) > 0$. Consequently, a liquid film with a thickness in a range where $\Pi(h) > 0$ can lower its free energy by becoming thicker in some areas while thinning in others, that is, by dewetting. When $\Pi(h) < 0$, wetting or spreading occurs.

The van der Waals interaction $w(r) \propto 1/r^6$ includes all intermolecular dipole–dipole, dipole–induced dipole, and induced dipole–induced dipole interactions. Performing a volume integral over all molecules present in the

Surface Tension Effects of Nanostructures,

Fig. 4 Illustration of EWOD. The external voltage is applied between a thin electrode (the Pt wire) and the bottom electrode (typically the indium-tin-oxide glass). Partially wetting liquid droplet in the absence of an applied voltage (**a**) and after the voltage is applied (**b**)



two half spaces bounding the film one finds a corresponding decay $\Pi(h) \sim A/6\pi h^3$ [11, 12], where the Hamaker constant A ($\sim 10^{-19}$ J) gives the amplitude of the interaction. In the “attractive” case, in which the layer tends to thin, $A < 0$. In the “repulsive” case, in which the layer tends to thicken, $A > 0$. Because of the disjoining pressure, there is a precursor film ahead of the nominal contact line of the liquid droplet [13]. The thickness of the precursor film can be defined by a molecular length [11, 12] $h_{PF} \sim \sqrt{A/6\pi\gamma_{LV}}$, which is on the order of several Å.

Key Research Findings

Elastic Models for the Nanostructures

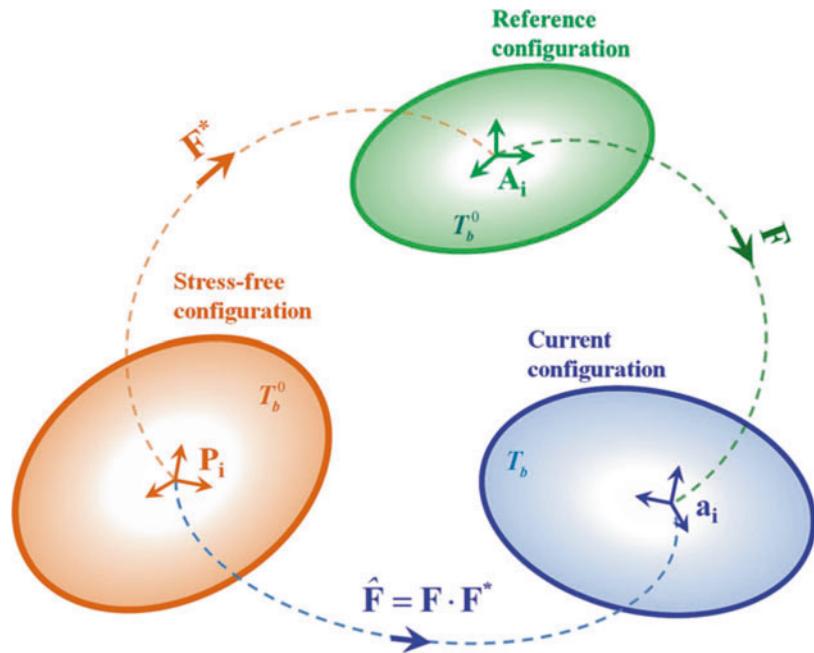
Gurtin and Murdoch established the theoretical framework of the surface elasticity under the classical theory of membrane [14]. Recently, studies [15] have shown that, even in the case of infinitesimal deformations, one should distinguish between the reference and the current configurations; otherwise the out-plane terms of surface displacement gradient, associated with the surface tension, may sometimes be overlooked in the Eulerian descriptions, particularly for curved and

rotated surfaces. By combining elastic models for surface and bulk, the size-dependent elastic and properties of nanomaterials have been investigated. Usually, in the absence of external mechanical or thermal loadings, the surfaces of a nanostructure will be subjected to residual surface stresses, and an elastic field in the bulk materials will be induced by such residual surface stresses induce from the point of view of equilibrium conditions. This self-equilibrium state without external loadings is usually chosen as the reference configuration, from which nanostructures will deform (see Fig. 5). That is to say, the bulk will deform from the residual stress states. However, in the prediction of elastic and thermoelastic properties of nanostructures, the elastic response of the bulk is usually described by classical Hooke’s law, in which the aforementioned residual stress was neglected in the existing literatures.

Considering a bulk material with the surface properties aforementioned, the surface tension would induce a stress field in the bulk. According to Young–Laplace equation, the surface tension will result in a nonclassical boundary condition. The boundary condition together with the equations of classical elasticity forms a coupled system of field equations to determine the stress

**Surface Tension Effects
of Nanostructures,**

Fig. 5 The choice of the reference configuration: the stressed state



distribution. To solve a problem considering the surface properties, the surface model and the bulk model are established separately and using the Young–Laplace equation to bridge the two models together.

Here, an example named surface elasticity would be used to show how to establish a model combined the surface and bulk together. Assuming the surface to be isotropic and homogeneous, the constitutive relations of the surface in the Lagrangian description can be written as [15]

$$\begin{aligned} \mathbf{S}_s &= \gamma_0^* \mathbf{I}_0 + (\gamma_0^* + \gamma_1^*) (\text{tr} \mathbf{E}_s) \mathbf{I}_0 \\ &\quad - \gamma_0^* (\bar{\nabla}_{os} \mathbf{u}_0) + \gamma_1 \mathbf{E}_s + \gamma_0^* \mathbf{F}_s^{(o)}, \end{aligned} \quad (10)$$

where \mathbf{S}_s is the first kind Piola–Kirchhoff stress of the surface, \mathbf{I}_0 is the identity tensor on the tangent planes of the surface in the reference configuration; the constants γ_0^* , γ_1^* , and γ_1 are the surface tension and the surface Lame moduli; \mathbf{E}_s , $\bar{\nabla}_{os} \mathbf{u}_0$, and $\mathbf{F}_s^{(o)}$ denote, respectively, the surface strain tensor, the in-plane part of surface displacement gradient and the out-plane term of surface deformation gradient. In view of the importance of the linearization of the general constitutive equations,

the linear elastic constitutive relations of the bulk with residual stresses can be written as follows:

$$\mathbf{S} = \mathbf{T}_R + \mathbf{u} \nabla \cdot \mathbf{T}_R + \lambda \text{tr}(\mathbf{E}) \mathbf{I} + 2\mu \mathbf{E}, \quad (11)$$

where \mathbf{S} is the first Piola–Kirchhoff stress, \mathbf{T}_R is the residual stress in the reference configuration, $\mathbf{u} \nabla$ is the displacement gradient calculated from the reference configuration, \mathbf{E} is the infinitesimal strain, and λ and μ are material elastic constants.

In the absence of external loading, surface tension will induce a compressive residual stress field in the bulk of the nanoplate and there may be self-equilibrium states which correspond to plate-self-buckling. The self-instability of nanoplates is investigated and the critical self-instability size of simplified supported rectangular nanoplates is proposed. The critical size for self-buckling is $b = \pi h \sqrt{(\alpha_A^{-2} + 1) Eh / [24\gamma_0^*(1 - v^2)]}$, where b and h are the width and thickness of the nanoplate, respectively; E and v denote the Young's modulus and Poisson's ratio, $\alpha_A = l/b$ is the aspect ratio [15].

Surface Tension Effects on the Mechanical Properties of Nanostructures

Many models are developed to relax one or some of the six above assumptions to extend Stoney's formula to a more generalized and realistic application. For example, the effects such as axial force, the damaged/nonideal interface effect and gradient stress, which violates one or several of the above assumptions, are analyzed and the extended/revised Stoney's formulas are given. Surface stress physically is a distributed one and this characteristic is not emphasized in many studies. The analysis by Zhang et al. shows that the Stoney's formula is obtained by assuming the influence of surface stress as a concentrated moment applied at the free end of a cantilever beam [16]. However, if the influence of surface stress is modeled as a distributed axial load and bending moment, the following nonlinear governing equation is obtained:

$$EI \frac{d^4 z}{dx^4} - \sigma b(L-x) \frac{d^4 z}{dx^2} + \sigma b \frac{dz}{dx} = 0, \quad (12)$$

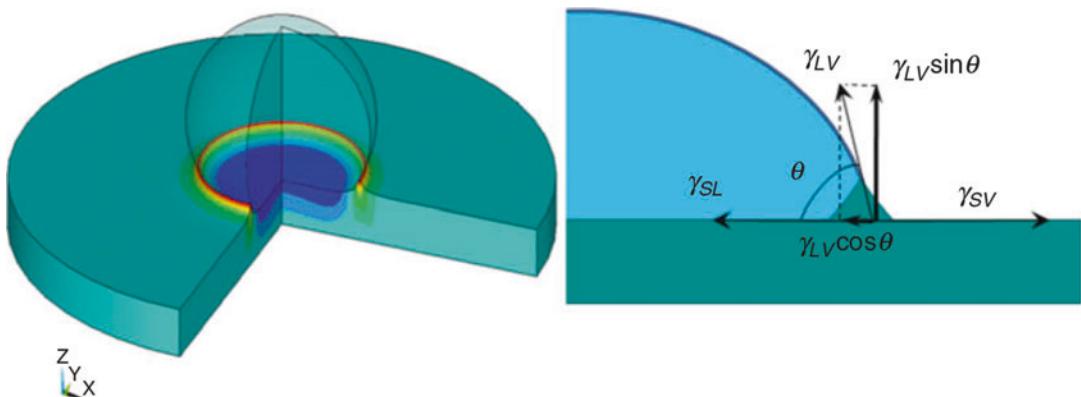
in which EI is the cantilever effective bending stiffness; b , L , and z are the beam width, length, and deflection, respectively. To solve the above nonlinear equation with the (given) boundary conditions at the two ends of the beam is a two-point-boundary value problem [16], which is rather difficult. The semi-analytical series solutions can more or less ease the difficulty of solving Eq. 12. One implication of the Stoney's formula is that because the beam curvature is constant, the beam deflection under a surface stress is an arc of a circle (or a parabola if the approximate curvature definition is used). There is no mechanism to guarantee such kind of deflection. In general the curvature of a beam under a surface stress is not a constant, which has been verified in the experiments.

The Stoney's formula has been used to explain recent experimental results, in which a hybrid device based on a microcantilever interfaced with bacteriorhodopsin (bR), undergoes controllable and reversible bending when the light-driven proton pump protein, bR, on the microcantilever surface is activated by visible light [17]. It should

be pointed that the Young's modulus of a nanostructure is size-dependent, that is, it would be enhanced or softened with decreasing the size of the nanostructure, which is generally attributed to the surface effects. Surface tension is one of the most important factors that cause the size effects of the Young's modulus of a nanostructure. The surface tension can be introduced into mechanical model via energy method. Using the relation of energy equilibrium, the effective elastic modulus of nanobeams are dependent on the surface tension [18].

Surface Tension Effects Induced the Deformation of Nanostructures

The classical Young's equation describes the equilibrium of forces in the direction parallel to the solid surface, while the vertical component of liquid–vapor interfacial tension is ignored. That is, there is a net force $\gamma_{LV} \sin \theta$ acting normal to the smooth solid surface at the solid–liquid–vapor contact line. Due to the unbalance force, there will be a surface deformation, as shown in Fig. 6. Indeed, several decades ago, surface deformation of semi-infinite solid was theoretically analyzed with the physical assumption that the liquid–vapor has a finite thickness (maybe at the order of tens nanometers) and the liquid–vapor interfacial tension acts uniformly in this region. Their research suggests there is a wetting ridge at the three-phase contact line. Later, Shanahan and Carré used height dimensional analysis to characterize the maximum at the order of γ_{LV}/G , where G is the shear modulus of solid [19]. For the material widely used at that time were very rigid (at the order of at least 100 GPa), such a deformation is too small to be considered. However, to meet with the rapid development of microelectromechanical systems (MEMS) and nanoelectromechanical systems (NEMS), polydimethylsiloxane (PDMS) is widely fabricated to channels or membrane, which has at least one dimension on the order of submillimeters or even nanometers. The surface deformation might no longer be neglected. Moreover, it should be noted that whether the theoretical solution for semi-infinite case can be extended to the case of thin flexible membrane. Recently, Yu and Zhao considered the



Surface Tension Effects of Nanostructures, Fig. 6 Sketch of deformation of PDMS membrane induced by a water droplet

deformation of thin elastic membrane induced by sessile droplet and gave a theoretical solution correspondingly [20]. There are two important conclusions. The first is that there exists a saturated membrane thickness at the order of millimeter, if the solid is thicker than this, it can be taken regard as semi-infinite; otherwise, it is better to consider the effect of membrane thickness. The second is that if the membrane has a very low Young's modulus (for example, on the order of MPa or much less), the effect of membrane thickness will become significant.

Apart from theoretical analysis, experimental investigations on surface deformation induced by droplet have also been reported. Because of the surface resolving ratio, it is difficult to get the detailed information of surface deformation at the contact line. Moreover, there might be a highly stressed zone near the contact line so that such a question should be studied further when necessary.

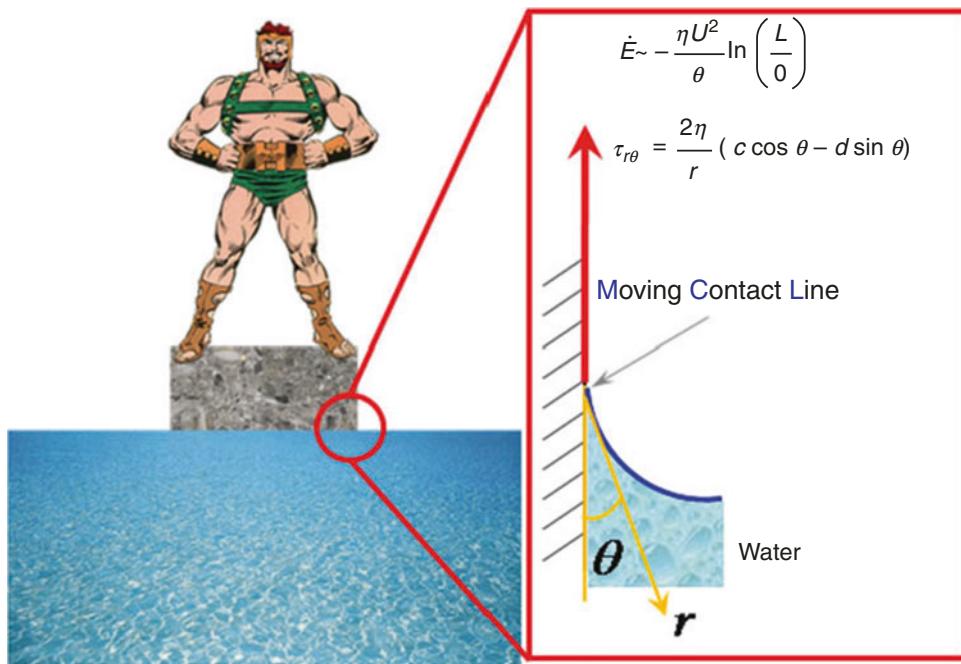
Surface Tension Effects of Wetting at the Three-Phase Contact Line

When considering the surface tension effects at the three-phase contact line, there is a famous paradox named Huh–Scriven paradox. It was first pointed out by Huh and Scriven [21] that there is a conflict between the moving contact line and the conventional no-slip boundary condition between a liquid and a solid. The interface

meets the solid boundary at some finite contact angle θ . Owing to the no-slip condition, the fluid at the bottom moves with constant velocity U and viscosity η , while the flux through the cross section is zero. The energy dissipation per unit time and unit length of the contact line is obtained, $\dot{x} \sim -\eta U^2 \ln(L/0)$, where L is an outer length scale like the radius of the spreading droplet. Stresses are unbounded at the contact line, and the force exerted by the liquid on the solid becomes infinite. The energy dissipation is logarithmically diverging, “not even Herakles could sink a solid” (Fig. 7)

In reality, dynamic wetting occurs at a finite rate with changes in the wetted area and liquid shape. These processes are thermodynamically irreversible and therefore dissipative. But, the energy dissipation is finite. There are two typical theories in identifying the effective channel of energy dissipation for small Capillary and Reynolds number. One of the two approaches is the hydrodynamic theory emphasizes energy dissipation caused by viscous flow within the wedge of liquid near the moving contact line. The other is the molecular kinetic theory emphasizes energy dissipation caused by attachment (or detachment) of fluid molecules to (or from) the solid surface [22].

The Huh–Scriven paradox is raised from four ideal assumptions summarized as: incompressible Newtonian fluid, smooth solid surface,



Surface Tension Effects of Nanostructures, Fig. 7 Huh and Scriven's paradox: not even Herakles could sink a solid

impenetrable liquid/solid interface, and no-slip boundary. Hence, the typical methods proposed to relieve the dynamical singularity near the contact line are the precursor film, surface roughness, diffuse interface, and nonlinear slip boundary, aiming the four assumptions respectively, as shown in Fig. 8.

in charging and discharging cycles hinders its applications. In recent years, nanostructured silicon is investigated as the material for electrode effectively circumvented the fragmentation, since surface tension effects of the nanostructured silicon on diffusion induced stresses would have much effect on the stress distribution [23].

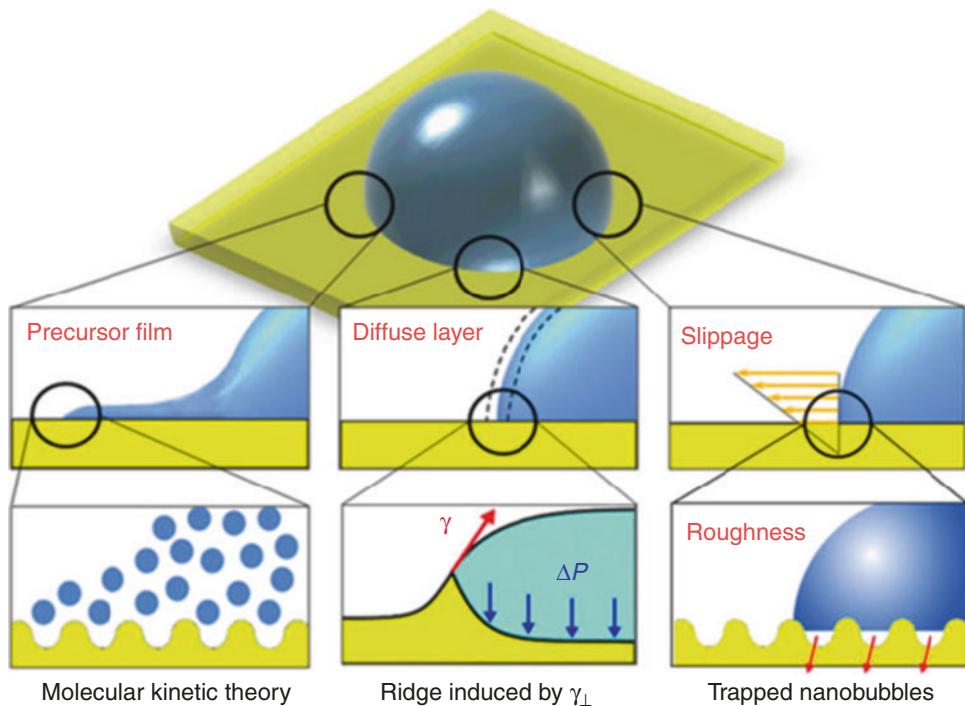
Examples of Application

Nanostructures of Silicon Used as the Anode Material of Lithium Ion Batteries

Rechargeable lithium ion batteries become the most suitable energy carrier for portable electro-equipments, electromobiles, and high performance computing, not only for the high-energy density and low cost, but also for the environmental needs for energy storage. Silicon is selected as a promising anode material of the lithium ion batteries due to the high-energy density (about 4,200 mAhg⁻¹). Nevertheless, a loss of electrical contact due to the fracture and crack of the bulk silicon induced by huge volume change (~400 %)

Surface Tension Effects Induced the Deflection of the Cantilever Sensor

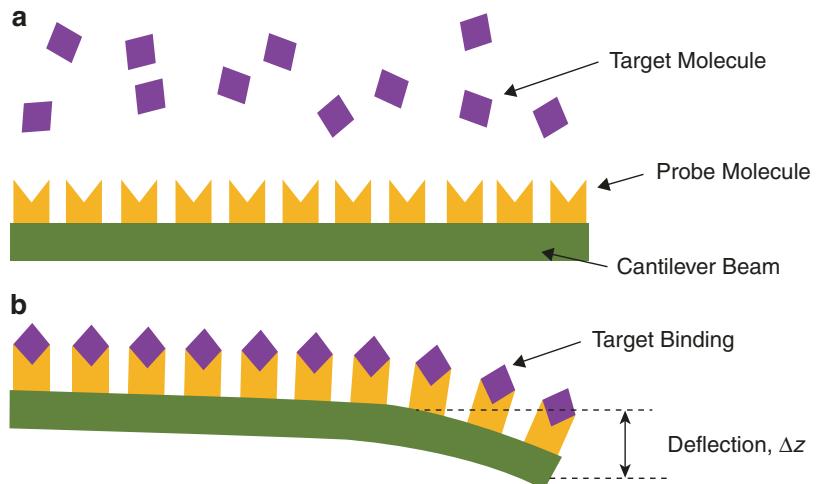
Surface stresses scale linearly with dimension and surface to volume ratio increases as micro/nano-structure scale decreases. Therefore, surface stress can be very important in microstructures in the size domain of MEMS/NEMS. Surface stress has been used as an effective molecular recognition mechanism. For a microcantilever used in a bioactuator, some biomaterials and biomolecules, which are immobilized on the microcantilever surface, are used to convert chemical energy into mechanical energy. When excitations such as DNA hybridization and receptor-ligand binding are applied, the microcantilever generates a



Surface Tension Effects of Nanostructures, Fig. 8 Possible mechanisms to solve the Huh and Scriven's paradox

Surface Tension Effects of Nanostructures,

Fig. 9 Schematic illustration of the deflection of a cantilever sensor due to the surface stress induced by surface tension effects [8, 17]. (a) The cantilever is functionalized on one side with the probe molecules. (b) After the target binding, the cantilever bends and the deflection can be measured



nanomechanical deflection response due to the surface tension effects, as shown in Fig. 9.

Surface Tension Effects Used in the Application of Electrowetting

One particularly promising application area for electrowetting is the manipulation of individual

droplets in digital microfluidic systems. Applications range from “lab-on-a-chip” devices to adjustable lenses and new kinds of electronic displays. Besides, electrowetting has been used for the application to displays showing video content since the switching speed is very high (only a few milliseconds) [10].

Summary

The surface tension effects become particularly predominant in nanostructures since the surface to volume ratio is sizable. In this entry, the surface tension effects that lead to the size-dependent mechanical properties of nanostructures are considered. Some key research findings related to this issue were listed and several examples of application were given, which are expected to be helpful to readers. Understanding and controlling these surface tension effects is a basic goal in the design and application of nanodevices.

Cross-References

- [Disjoining Pressure and Capillary Adhesion](#)
- [Electrowetting](#)
- [Nanoscale Properties of Solid–Liquid Interfaces](#)
- [Surface Energy and Chemical Potential at Nanoscale](#)
- [Wetting Transitions](#)

References

1. Bhushan, B. (ed.): *Handbook of Nanotechnology*. Springer, New York (2010)
2. Gibbs, J.W.: On the equilibrium of heterogeneous substances. In: Gibbs, J.W. (ed.) *The Scientific Papers of J. Willard Gibbs*. volume 1: Thermodynamics, pp. 55–353. Dover, New York (1961)
3. Cahn, J.W., Hilliard, J.E.: Free energy of a nonuniform system. I. Interfacial free energy. *J. Chem. Phys.* **28**, 258–267 (1957)
4. Lu, W., Suo, Z.: Dynamics of nanoscale pattern formation of an epitaxial monolayer. *J. Mech. Phys. Solids* **49**, 1937–1950 (2001)
5. Butt, H.J., Graf, K., Kappl, M.: *Physics and Chemistry of Interface*. Wiley-VCH, Weinheim (2003)
6. Shuttleworth, R.: The surface tension of solids. *Proc. Phys. Soc. A* **63**, 444–457 (1950)
7. Herring, C.: Surface tension as a motivation for sintering. In: Kingston, W.E. (ed.) *The Physics of Powder Metallurgy*, pp. 143–179. McGraw Hill, New York (1951)
8. Fritz, J., Baller, M.K., Lang, H.P., Rothuizen, H., Vettiger, P., Meyer, E., Guntherodt, H.J., Gerber, C., Gimzewski, J.K.: Translating biomolecular recognition into nanomechanics. *Science* **288**, 316–318 (2000)
9. Stoney, G.: The tension of metallic films deposited by electrolysis. *Proc. R. Soc. Lond. A* **82**, 172–175 (1909)
10. Berthier, J.: *Microdrops and Digital Microfluidics*. William Andrew, Norwich (2008)
11. De Gennes, P.G.: Wetting: statics and dynamics. *Rev. Mod. Phys.* **57**, 827–863 (1985)
12. De Gennes, P.G., Brochard-Wyart, F., Quere, D.: *Capillarity and Wetting Phenomena*. Springer, New York (2004)
13. Yuan, Q.Z., Zhao, Y.P.: Precursor film in dynamic wetting, electrowetting and electro-elasto-capillarity. *Phys. Rev. Lett.* **104**, 246101 (2010)
14. Gurtin, M.E., Murdoch, A.I.: A continuum theory of elastic material surfaces. *Arch. Ration. Mech. Anal.* **57**, 291–323 (1975)
15. Wang, Z.Q., Zhao, Y.P., Huang, Z.P.: The effects of surface tension on the elastic properties of nano structures. *Int. J. Eng. Sci.* **48**, 140–150 (2010)
16. Zhang, Y., Ren, Q., Zhao, Y.P.: Modelling analysis of surface stress on a rectangular cantilever beam. *J. Phys. D Appl. Phys.* **37**, 2140–2145 (2004)
17. Ren, Q., Zhao, Y.P.: A nanomechanical device based on light-driven proton pumps. *Nanotechnology* **17**, 1778–1785 (2006)
18. Guo, J.G., Zhao, Y.P.: The size-dependent bending elastic properties of nanobeams with surface effects. *Nanotechnology* **18**, 295701 (2007)
19. Shanahan, M.E.R., Carré, A.: Nanometric solid deformation of soft materials in capillary phenomena. In: Rosoff, M. (ed.) *Nano-Surface Chemistry*. CRC Press, New York (2001)
20. Yu, Y.S., Zhao, Y.P.: Elastic deformation of soft membrane with finite thickness induced by a sessile liquid droplet. *J. Colloid Interface Sci.* **339**, 489–494 (2009)
21. Huh, C., Scriven, L.: Hydrodynamic model of steady movement of a solid/liquid/fluid contact line. *J. Colloid Interface Sci.* **35**, 85–101 (1971)
22. Wang, F.C., Zhao, Y.P.: Slip boundary conditions based on molecular kinetic theory: the critical shear stress and the energy dissipation at the liquid–solid interface. *Soft Matter* **7**, 8628–8634 (2011)
23. Cheng, Y.T., Verbrugge, M.W.: Evolution of stress within a spherical insertion electrode particle under potentiostatic and galvanostatic operation. *J. Power Sources* **190**, 453–460 (2009)

S

Surface Tension–Driven Flow

- [Capillary Flow](#)

Surface Tension–Powered Self-Assembly

- [Capillary Origami](#)

Surface-Enhanced Raman Scattering for Imaging Biological Cells

Nicolas Pavillon¹, Katsumasa Fujita² and Nicholas Smith¹

¹Biophotonics Lab, Immunology Frontier Research Center (IFReC), Osaka University, Osaka, Japan

²Department of Applied Physics, Osaka University, Osaka, Japan

Definitions

Nanoparticles, as a term, generally refer to any type of particle that is composed of metallic, semiconductor, or even biological material. The particles sizes that fall under the term usually range from around 1 to 100 nm in diameter. Very small size nanoparticles, where the number and arrangement of atoms are well determined, are also termed nanoclusters (e.g., a 22-atom gold nanoparticle is usually referred to as a nanocluster). In the current article, the term “nanoparticles” is used to refer to 50 nm diameter gold nanoparticles, unless otherwise noted. Surface-enhanced Raman scattering (SERS) is a term that refers to the amplification of the wavelength shift in scattered light where the shift is caused by vibrations in a sample (similar to a Doppler shift), and the amplification is caused by the presence of a surface. “Biological cells” usually refer to living cells, either in tissue or cultured on a substrate. Whether mammalian or not, cells are the building blocks of organisms and are therefore primary targets of interest for these emerging nanoparticle-based analytic methods. SERS imaging in biological cells is then the use of these particles in cells and the methods of attempting to use the resulting spatial and spectral information for analyzing the cell state and in particular the molecular conditions around the particle. This can be either on the cell membrane surface or inside the cell if the particles can be moved to positions of interest in the cell.

Overview

Nanoparticle types suitable for biological SERS measurements: For biological applications where SERS is a key interest, the particle material is limited to compositions that have plasmonic properties over suitable light wavelength ranges. Usually the toxicity of nanoparticles themselves and/or the phototoxicity of the light which excites the SERS effect needs to be reduced to as low a level as possible [1]. Practically, this means that most applications use inert gold as the basis for the nanoparticles, since gold is relatively chemically inert [2] while still exhibiting significant plasmonic enhancement for suitable particle sizes. Even though gold is a precious metal, the cost is usually not a significant factor in the choice of metal; the amount of gold used in typical SERS imaging experiments using cultured cells in a dish is on the order of a few US dollars or less.

For a given metal, the optimum nanoparticle size is then also bounded by a number of physical limitations: the particle must be large enough to undergo plasmonic enhancement at excitation wavelengths compatible with intracellular experiments, and it must be small enough to allow uptake or injection into a cell. It must be of a size that does not significantly interfere with the homeostasis of the cell. Very small nanoparticles (~ 2 nm) are thought to intercalate with nucleic acids, while very large particles will either be excluded by the cell or may significantly disrupt normal cellular function if taken up by a cell. Taken together, these constraints usually result in the choice of metal nanoparticles of approximately 50 nm diameter and spheroidal shape, with gold as the metal. These can be synthesized in a number of different ways; however, since they are commercially available, explicit treatment of synthesis methods is not covered in this current work. There are a number of points to be aware of in the purchase and/or synthesis of such gold nanoparticles. For example, the surface coating can significantly change the uptake, aggregation, toxicity, and surface charge of the nanoparticles. Some nanoparticles in literature reports are

referred to as “naked nanoparticles,” but this usually refers to a minimal surface coating that is not thought to influence the properties of the nanoparticles significantly. Truly naked (i.e., without any surface coating) nanoparticles are difficult to prepare as a colloidal solution since aggregation would make the solution unstable.

Non-spheroidal Nanoparticles

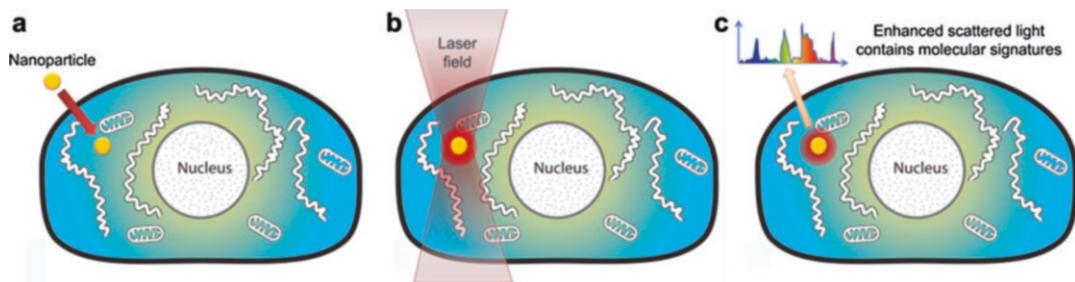
The plasmonic properties of metal nanoparticles depend on the type of metal and on the size and geometry of the particle. The most widely used particle shapes for intracellular measurements are 50 nm spherical gold particles; however, a very large range of complex particles shapes can be formed by so-called “bottom-up” methods [3]. In contrast to cutting shapes out by electron beam or other patterned methods, the shape is formed in bottom-up methods by controlling the fabrication parameters so that the desired shape emerges as a result of the fabrication reaction. These wide-ranging shaped nanoparticles have a number of applications. One commonly used technique is to use nanoparticles that are approximately cylindrical in shape; this then gives a long and short axis. Due to the plasmonic resonances, there are typically two resonance peaks associated with the cylindrical geometry, and the particles can be readily excited by wavelengths between the visible and near-infrared range. Another type of particle geometry involves using nanoshells where the outside shell of the particle is made of a plasmonic material (often gold) and the inner core can be made of a variety of different materials. This then allows much more sophisticated tuning of the plasmonic properties. Even for spherical-shaped nanoshells, the breaking of symmetry (whereby the core center is not located at the center of the particle) allows significant modification or properties.

Plasmonic enhancement is best treated in dedicated articles, which are numerous, but are only recently entering the public awareness [4]. For the purposes of this current work, nanoparticle

plasmonic enhancement can be briefly stated to be the interaction (i.e., the coupling) between the “plasma” or electrons in the metal and an applied external light field. The electrons are affected by the external field and move in response as waves on the surface of the metal, which leads to the resonant properties of the effect and thereby the dependencies on both particle geometry and the wavelength of the external field.

Practical Use of Nanoparticles for Cellular Imaging

Even the use of the word “imaging” requires some clarification since particles are usually only present in low numbers and in relatively few regions of the cell. Starting from the idea of a cell without nanoparticles (Fig. 1a), and by some means the particles enter the cell, so that if a laser of the right wavelength is irradiated onto the region containing the nanoparticle (Fig. 1b), a local-enhancement occurs, and the enhanced Raman signal is then read out from the cell (Fig. 1c). Getting nanoparticles into cells is in itself often a challenge, and a number of different options are available to the experimentalist to induce the uptake. For some cell types (most notably macrophages), uptake of general debris is one of the biological functions of the cell, and for these types, nanoparticles are relatively quickly taken up. For other types, such as HeLa cells, which do not have specific uptake roles in the body, nanoparticles will still enter, but at a much lower rate. This is not necessarily a limitation but it should be taken into consideration when designing experiments. If nanoparticle-enhanced SERS signals are an integral part of the measurements, then long incubation times (as a rough estimate, on the order of 24 h but with wide variance) may be required to have a sufficient number of particles inside the cell in order to achieve meaningful results. Additionally, by modifying nanoparticle surface properties, or the culturing conditions, these uptake rates can be significantly changed.



Surface-Enhanced Raman Scattering for Imaging Biological Cells, Fig. 1 Overview of nanoparticle-cell measurements by surface-enhanced Raman scattering: Nanoparticles enter an otherwise normal cell (a), where

an external laser field then excites a strong local field around the nanoparticle (b). With the possibility of additional chemical enhancement, the nanoparticle can be used to report the local Raman molecular signature in the cell (c)

Targeted Versus Tagged Versus Naked Nanoparticles

A major distinction between different types of intracellular SERS measurements is whether nanoparticles are modified to target them to particular intracellular molecules, tagged so that they contain their own labels [5] or left relatively untreated to attempt unbiased measurement of endogenous intracellular molecules. For targeting, the particles can be conjugated with a compound that binds with an intracellular target molecule. Tagging can be achieved by using a layer of Raman-active material which results in enhanced emission, often with an additional layer of inert material over the Raman-active material. These tagged nanoparticles can then be additionally targeted to intracellular molecules by conjugation with an appropriate molecule. Even in the case of the so-called “naked” nanoparticles, where particles are expected to enter the cell, provide a base for surface enhancement, and thereby act as a means to report the intracellular milieu, the idea of an unbiased measurement of intracellular molecules using gold nanoparticles to probe them is not strictly accurate. The reason is that there are a number of effects at play in the surface enhancement of molecules. Two main contributing roles are played by two fundamentally different physical mechanisms. Electromagnetic enhancement of the field occurs near the surface and will enhance both the incident field and the emitted SERS field, while a chemical enhancement also occurs due to chemical interactions between the surface and the

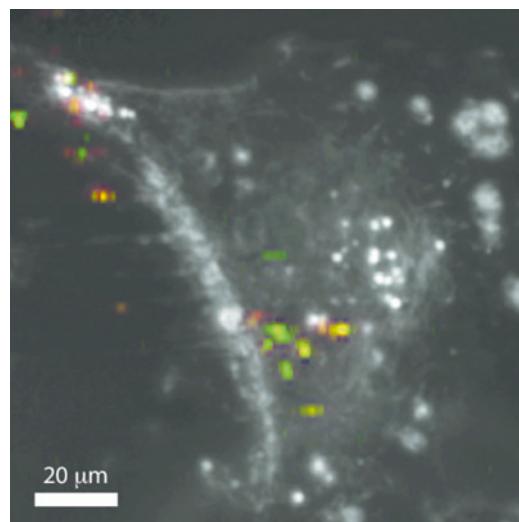
molecule, increasing the probability of generating Raman-scattered photons. The exact role of these two modes in SERS was debated for a number of years in the literature, but what is important for the present discussion is not so much the relative role of each but rather that they both exist, and therefore two different molecules in the same enhanced field will experience a different probability of generating SERS, leading to a description of molecules being “SERS-active” or not. Other effects coming into play include the fact that the same molecule can interact with a surface in different ways, with different orientations, and therefore produce a difference not only in the intensity of the SERS signal but also in the nature of the signal provided. Some Raman bands will not be active in SERS, some will be enhanced, and some will show significant dependence on the exact nature of the chemical interaction with the substrate, incident field polarization, fluctuations in time with diffusion, and other parameters. This leads to issues which must not only be dealt with at the time of interpreting data but must also be considered when designing the experiment. As a rough example, lipids tend to be poor molecules for intracellular observation by SERS. If lipids were a key target, then conjugated nanoparticles targeted specifically for lipids would be a relevant choice (although practically conjugation possibilities for lipids are limited). If, instead, it was desired to attempt to measure the endogenous molecules responsible for endocytosis and subsequent transport, naked nanoparticles would be a reasonable choice since they assume nothing

about the molecules involved and may instead report the local molecules which transport the nanoparticle (subject to the inherent bias resulting from the difference in SERS activity for different molecules).

Measuring SERS

The next issue in intracellular SERS measurements would be the choice of an appropriate wavelength for the laser which excites the SERS signals. While the use of the nanoparticle imparts some plasmonic effect which is needed to boost the signal to a detectable level, it is often not straightforward to choose the laser wavelength. Intuitively, matching the laser wavelength to the plasmon resonance would seem to be an ideal choice, but practically, there are reasons to avoid this. When we are using gold nanoparticles of nominally 50 nm diameter (for reasons outlined above), the plasmon resonance is typically around 532 nm. Choosing an excitation laser wavelength of 532 nm (which happens to be one of the most common laser wavelengths) tends to excite luminescence from the gold, which is a broadband, red-shifted light output that easily overwhelms the relatively weak SERS signal. The excitation by 532 nm can give a precise image of the locations of nanoparticles due to this luminescence, but we usually want to observe the SERS signals. This then leads to a trade-off: at 532 nm we will have strong plasmonic enhancement but overwhelming luminescence. At longer wavelengths between say 640 and 800 nm, we will have significantly less luminescence but also reduced SERS signals. It is still practically beneficial to choose a wavelength between approximately 600 and 800 nm, and most SERS experiments inside cells use this wavelength range.

Detection is a complex issue for this type of experiment. Before determining how detection can be achieved, we need to consider how the particles might be located in the cell. Particles do not cover the entire cell, and SERS will only be observed in regions where particles exist (and often specifically at hotspots between particles). If we image the entire field of view using, for



Surface-Enhanced Raman Scattering for Imaging Biological Cells, Fig. 2 Image of a HeLa cell with gold nanoparticles. In grayscale is shown a dark-field image taken before SERS measurement. The red channel shows the position of GNPs within the cell, and the green one presents the quality map for the spectra detection (See Pavillon et al. [6] for details)

example, the excitation wavelength of 780 nm, then the resulting image would be detected through the spectrometer (details clarified below) and the SERS signals would appear in locations where particles existed. This is outlined in Fig. 2. Here, the cell was first imaged using a dark-field imaging mode which highlights strongly scattering objects (including nanoparticles as well as endogenous scattering molecules such as lipid vesicles and cellular structure), analogous to the way that dust in the air is often visible when a beam of early morning light strikes into a room through a curtain. Overlaid on the dark-field imaging mode is a color-mapped red channel that corresponds to the locations of nanoparticle in the cell, as determined from luminescence measurements, and also a green channel which shows “high-quality” SERS signals detected from the SERS mode (we will clarify what “high quality” means below). In areas that have both green and red, i.e., where the locations of particles correspond perfectly to the locations of “high-quality” SERS signals, the pixels should appear yellow. Note that very few of the pixels in the color

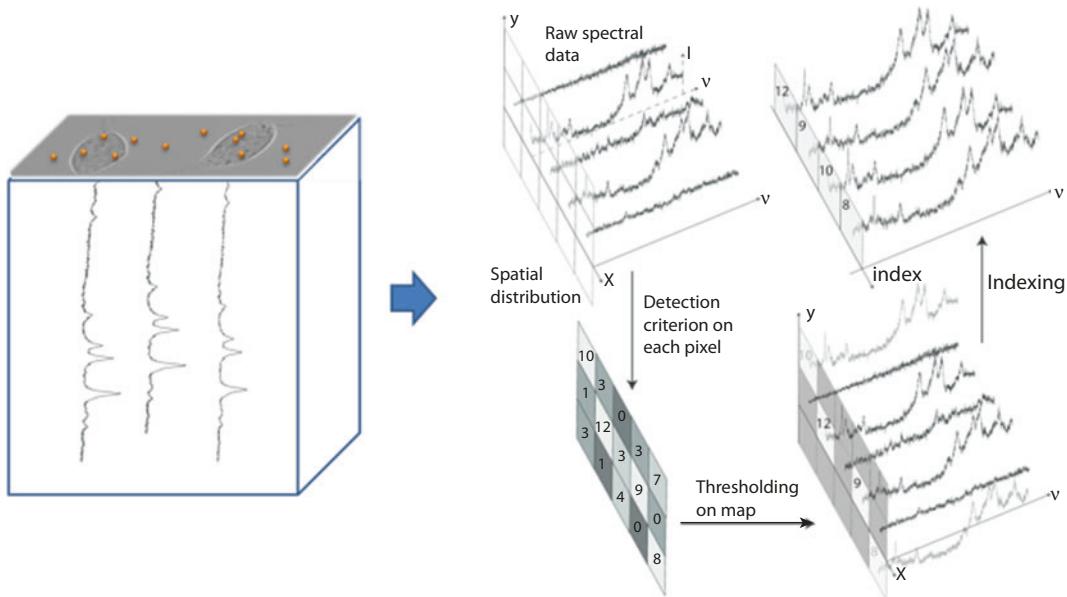
overlay actually appear yellow. This gives us a strong hint that the detection of spatial arrays of SERS is actually a significantly complex issue that must be carefully considered before acquiring SERS measurements from a cell.

For labeled nanoparticles, the SERS measurements can be carried out in a manner analogous to fluorescence measurements, with the exception that rather than a fluorescence excitation process, the SERS process is generated by the excitation laser and emission is detected. With a full spectrum expected from each pixel, the simplest method is to raster scan the excitation beam and then to pass the emission light through a spectrometer, acquiring the emission spectrum for each point as rapidly as possible. The overall imaging rate will be limited by this spectral readout time for each spatial pixel. Since modern spectrometers are available with “imaging” capability and with a 2D detector, it is possible to excite an entire line of the sample and, for each point on the line, to resolve the spectra at the detector plane of the spectrometer. This means each full-frame readout of the 2D detector at the spectrometer provides one spatial line in the sample plus the corresponding spectral information for each pixel in that line. Such line-scanning SERS methodology is now commercially implemented and available, making frame rates of roughly one minute per frame for a frame size of 400 by 400 pixels, although it should be noted the frame rates will depend very significantly on a number of factors, including the sample type, exposure time, excitation wavelength, power, etc.

Obtaining Meaning from Large SERS Datasets

For a 2D image of cells containing nanoparticles, a key issue is that before measuring we often don’t know (a) where the nanoparticles are in the cell and (b) whether those nanoparticles will provide any SERS signals. There are several approaches we can choose to accommodate this lack of knowledge. For tagged or labeled nanoparticles, the use of a Raman reporter can make the emission strong enough that the particles effectively behave

like fluorescent tags (albeit with the possibility of multiplexing due to the discrimination ability of the additional spectral dimension). For experiments where the SERS signals result from molecules outside the nanoparticle, there is usually significant variation in the signals, depending on the local environment in the cell, as well as fluctuations in time. If we simply add nanoparticles to a cell culture, wait for the requisite time for particle uptake, and measure the resulting spatial distribution by SERS imaging, the resulting dataset is a 3-dimensional stack that will contain some SERS signals at some locations. Along with the SERS data, background signals including fluorescence, cosmic rays, and surface contamination of the nanoparticles will often appear. This means that after acquiring the SERS dataset, it is still challenging to determine which data is relevant for the SERS experiment. While nanoparticles are required to generate SERS signals, the SERS signals do not appear at all nanoparticle locations, due to a number of effects (such as gaps between particles, differences in molecule-surface interactions, and others described above). This is demonstrated in Fig. 3. The resulting dataset will be composed of mostly non-SERS-related pixels. Classification of SERS spectra can be done by algorithm, but careful choice and testing of the algorithm on known datasets should be done before application to new or unknown datasets. For example, a simple algorithm to separate SERS data from non-SERS data would be to select the pixels that contain the most spectral energy, under the assumption that the surface-enhanced regions should be brighter than the background. This is often not the case though for real data and is further complicated by the fact that the surface enhancement also applies to background signals and fluorescence. An algorithm that then classifies SERS signals based on spectral peaks, spectral energy and other properties specific to SERS can be developed to recognize SERS features. Pixels ranked high in the SERS detection algorithm can be loosely described as having “high-quality” SERS spectra. Once this information is known and the locations of those spectra are known, higher-order analysis can be applied whereby spectra are further sorted into clusters of similar



Surface-Enhanced Raman Scattering for Imaging Biological Cells, Fig. 3 HeLa cells incubated with gold nanoparticles, showing how nanoparticle distribution does not match detected SERS distributions. The large datasets resulting from imaging SERS contain a large number of pixels with background signals which do not contribute to

the SERS information of interest. One approach to simplify the data is to design an algorithm to remove pixels which have a spectra that do not have SERS features, by applying a detection criteria based on known properties of SERS spectra and thresholding data that does not meet the criteria [6]

spectra, and questions such as “do all cells show the same variations in spectra?,” “are all spectra completely randomly distributed within a cell?,” or “are there correlations between variance in spectra and cell type or function?” can all be asked and answered from the data analysis [6].

A Tracking Approach to Nanoparticle Dynamics and SERS Measurement

The method described above measures the entire field of view and then subjects the data to analysis using customized algorithms that recognize SERS features. An alternative to measuring the entire field is to instead track the location of a single particle and follow its trajectory with the SERS excitation laser beam [7]. This allows the possibility of discrimination of very rapid changes in SERS signal in time. If the SERS excitation laser follows the particle, then the SERS detection scheme essentially reduces to 1-dimensional detection (the point is read out as a 1D spectral array). The minimum

exposure time can then be reduced to an interval as short as several tens of milliseconds. SERS signals usually contain a significant amount of inherent temporal fluctuations, which can derive from the combination of molecular movement and the very high sensitivity of the SERS effect. Such dynamics are a source of information, particularly inside a cell where molecular motion may result from simple Brownian motion or by directed intracellular transport. By reducing the spatial dimensions to a point, i.e., with the loss of spatial information, there is the possibility of a significant gain in temporal information. Even with the SERS detection only occurring from a single point in the sample, the technique can still be considered a type of imaging: in order to track and follow the particle motion with the SERS laser, a secondary imaging mode is needed which provides an image of the particle location. This mode can be a simple imaging mode, such as dark-field microscopy, and due to the a priori knowledge of the nanoparticle shape, a centroid detection algorithm can be employed to estimate the center of the particle to a level much

finer than the overall microscope resolution. The particle trajectory can then be tracked with a resolution of tens of nanometers, which is a scale of particular importance to cell biology and active transport. The combination of the particle trajectory information with the SERS feedback allows the additional possibility of correlating the two sets of information. This is important because it allows the determination of which types of SERS signals correspond with which types of particle motion. The typical SERS spectra for Brownian motion can be distinguished from typical SERS spectra for directed transport within the cell, and this is a good starting point to use SERS-based analysis to help discover molecules involved in the cellular transport mechanism.

Other Approaches to SERS Measurements of Cells

With the possibilities of tracking single molecules, high resolution, and making new discoveries in the cell, it is not surprising that a number of approaches to SERS measurements in cells have been developed, not limited only to the ones described above. For example, nanoparticles can be created inside the cells by laser induced photoreduction, and then used to read out molecular signatures [8]. Additionally, SERS is not limited only to the interior of cell samples. It is often easier, in practice, to use SERS as a tool to probe membrane-based targets, since it does not require cellular uptake. Incubation of gold nanoparticles with cultured cells results in settling of some portion of the nanoparticles on the cell surface within several minutes. These can then be used to attempt to detect surface receptors or uptake pathways. It is also relatively common to use a SERS-active material as a substrate for cultured cells. This allows the investigation of the basal cell membrane, with a particular affinity for cell adhesion molecules, since they are the regions actually in contact with the substrate. A substrate approach has a number of key features: the detection range in the axial direction (i.e., away from the substrate) is extremely short and on the order of a few tens of nanometers. Additionally, the

substrate itself should be relatively stable so that cell movements should be movements across the substrate and not dynamics of the substrate itself. This can simplify the analysis of SERS signals and negates the requirement to account for changes of the SERS sites themselves which can complicate analysis in intracellular nanoparticle-based SERS. It is also possible to fabricate gold nanoparticles that can be used for SERS directly in targeted locations inside a cell. This approach uses focused laser light to generate a photoreduction reaction of gold ions and form gold nanoparticles at the laser focus. This makes the task of finding the SERS locations much simpler since they appear at the laser focus itself. Currently, SERS measurements in cells are already demonstrated to be a method by which unique information on cell content can be determined. It is still an area under intense development, however, and very significant advances are to be expected over the next decade.

Cross-Reference

► Plasmonics

References

1. Sutariya, V.B., Yashwant, P.: *Biointeractions of Nanomaterials*. CRC Press, Boca Raton (2014)
2. Mohr, F. (ed.): *Gold Chemistry: Applications and Future Directions in the Life Sciences*. Wiley-VCH, Weinheim (2009)
3. Sau, T.K., Rogach, A.L. (eds.): *Complex-Shaped Metal Nanoparticles*. Wiley-VCH, Weinheim (2012)
4. Atwater, H.A.: The promise of plasmonics. *Sci. Am.* **296**, 56–62 (2007)
5. Chau, L.-K., Chang, H.-T.: From Bioimaging to Biosensors: Noble Metal Nanoparticles in Biodetection. Pan Stanford, Singapore (2013)
6. Pavillon, N., Bando, K., Fujita, K., Smith, N.I.: Feature-based recognition of surface-enhanced Raman spectra for biological targets. *J. Biophotonics* **6**, 587–597 (2013)
7. Ando, J., Fujita, K., Smith, N.I., Kawata, K.: Dynamic SERS imaging of cellular transport pathways with endocytosed gold nanoparticles. *Nano Lett.* **11**, 5344–5348 (2011)
8. Smith, N.I., Mochizuki, K., Niioka, H., Ichikawa, S., Pavillon, N., Hobro, A.J., Ando, J., Fujita, K., Kumagai, Y.: Laser-targeted photofabrication of gold nanoparticles inside cells. *Nat. Comm.* **5**, 5144 (2014)

Surface-Modified Microfluidics and Nanofluidics

Kaushik K. Rangharajan and Shaurya Prakash

Department of Mechanical and Aerospace Engineering, The Ohio State University, Columbus, OH, USA

Definition

Fluid phenomena in micrometer (1–100 μm) or nanometer (1–100 nm)-sized channels that are governed by coupled principles from fluid mechanics, surface chemistry, electrochemistry, and electrostatics are generally referred to as microfluidic and nanofluidic phenomena [1]. The smaller length scales compared to traditional fluid mechanics provide several new and interesting phenomena due to significant enhancement of the surface-area-to-volume ratio, which makes surface-mediated flows important to the overall field of microfluidics and nanofluidics.

Introduction to Surfaces in Microfluidics and Nanofluidics

Microfluidic and nanofluidic devices and systems are characterized by high surface-area-to-volume (SA/V) ratio. For example, a rectangular cross-section channel with 100 μm width, 100 nm depth, and 1 cm length will have a SA/V ratio on the order of 10^6 m^{-1} . In fact, operational devices incorporating components with SA/V ratio on the order of 10^9 m^{-1} have already been reported [2]. Consequently, the influence of device and system walls in affecting phenomena within confined spaces can no longer be ignored since the walls (i.e., surfaces) of the devices and systems interact extensively and directly with the species contained within these devices. Therefore, the surface of interest here is defined as an interface or thin region in space (often only a few nm in physical extent) that influences transport and reaction phenomena in its vicinity. In this surface-region properties such as chemical composition,

refractive index, mechanical strength, conductivity, and charge can significantly differ from the bulk material of the underlying substrate.

Scaling of Surface Forces in Microchannels and Nanochannels

One non-dimensional parameter often used to characterize flow regimes is the ratio of the inertial forces transferred by the velocity (or momentum) of the fluid to the viscous (or frictional forces), which is the Reynolds number, Re , and is given by

$$Re = \frac{\rho VL_c}{\mu}, \quad (1)$$

where, ρ is the fluid density, μ is the fluid viscosity, and V is the fluid velocity.

As seen from Eq. 1, Re is directly proportional to the characteristic length, L_c of the flow. With L_c decreasing (and consequently SA/V increasing), the inertial forces decrease, and as a result the Re becomes smaller. The direct consequence is that for given flow parameters of fixed velocity and fluid type, a decreasing L_c implies increasing influence of viscous forces in contrast to the inertial forces.

In most microfluidics and nanofluidics applications, a particle (ion, colloid, biomolecule, AFM probe tip, etc.) will interact with the channel walls. Considering three important forces – electrostatic or Coulombic (F_{el}), van der Waals (F_{vdw}), and hydrodynamic forces (F_h) – will provide a better insight between surface-particle interactions and subsequent phenomena. For an infinite flat surface separated by a distance D from a particle of radius R , F_{el} is given by [3]

$$F_{el} = -\frac{2\pi RL_D}{\epsilon\epsilon_0} [2\sigma_S\sigma_P \exp(-D/\lambda_D) + (\sigma_s^2 + \sigma_p^2) \exp(-2D/\lambda_D)] \quad (2)$$

where, λ_D is the Debye length and σ_s and σ_p are the surface charge densities of surface and particle, respectively. F_{vdw} is estimated by [3]

$$F_{vdW} = \frac{A_H R}{6D^2}, \quad (3)$$

where A_H is the Hamaker constant, which serves as an indicator of the interaction between the particle and the sample surface, and F_h has been approximated by [3] the expression

$$F_h = b_s V = -f^* \frac{6\pi\mu R^2 V}{D}, \quad (4)$$

where V is the particle velocity, b_s is the hydrodynamic damping coefficient, and μ is the viscosity of the fluid. The coefficient f^* is related to the boundary slip properties. If the no-slip assumption holds true at the solid/water interface, then $f^* = 1$, and if slip exists, then $f^* < 1$. As an aside, it should be noted that Eq. 4 is often used in literature for determining slip lengths using a colloidal AFM probe tip. As the operational length scales (here, characteristic length scale is separation distance, D) decrease, it can be seen from Eqs. 2 to 4 that contributions for F_{el} display an exponential dependence, F_{vdw} display an inverse quadratic dependence, and F_h display an inverse linear dependence on the separation distance. Therefore, consideration of each force term at the microscale, and even more so at the nanoscale, is essential toward completely describing importance of the wall-species interactions.

Most microfluidic and nanofluidic devices with liquid flows are driven by electric fields and fall under the category of electrokinetic flows. The main reason for using applied voltages to generate electric fields for driving flows is the scaling of pressure forces. For example, for a circular nanochannel with laminar, incompressible, Poiseuille flow with water as the working fluid, the necessary pressure drop across a 100 μm long channel which is 1 nm in diameter for only an attoliter (10^{-18} m^3) per second would be greater than 3 GPa, which is impractical for any device. Electrokinetic flows can sustain higher flow rates through nanometer channels without excessive pressures [2].

Methods for Surface Modification

Surface modification methods can be divided in two broad categories: physical and chemical methods [4]. Physical methods, in most cases, do not change the chemical composition of the surface but may change the surface roughness, grain sizes and grain boundaries, and faceting. Physical methods often use lasers, plasmas, temperature changes, ion beams, ball milling, and polishing or grinding, to alter the surface state of a material of interest. While the main intent with physical modification methods is to not alter the chemical composition of the material, in some cases, physical surface modification methods can lead to changes in the chemical composition of the surface due to removal or addition of material or chemical reactions on surfaces, as in the case of selective or ion-beam sputtering, or by selective cross-linking in the presence of plasmas. Temperature gradients and thermal treatments have been used to change surface roughness, grain sizes, and grain boundaries and create nanoscale features, facets, textures, and nanoparticles on ceramics, metals, polymers, and semiconductors. Thermal treatments in the presence of gases such as oxygen or water vapor can cause creation of steps or induce other forms of nanostructures. For example, as was discussed in a recent review [4], crystalline $\alpha\text{-Al}_2\text{O}_3$ surfaces heat treated at 1,500 °C in Ar/O₂ and H₂/He/O₂ led to step formation and roughening as quantified through AFM topology images.

Chemical methods introduce a change in the chemical composition at the surface by introducing chemical properties (e.g., surface charge density or surface energy) different from the bulk material. Among the methods for chemical modification formation of surface layers, either covalently bonded or physisorbed, has been most common. Other chemical methods include treatment with UV light and reactive plasmas. Modification schemes are governed by a wide range of parameters including sample type (polymers, metals, ceramics, etc.), stability to treatment conditions (e.g., thermal or structural), and eventual applications. For example, polymeric surfaces are often modified by photochemical methods of

which UV irradiation in air, other reactive atmospheres such as ozone combined with lasers, and grafting surface layers are fairly common. The use of reactive plasmas has been gaining popularity due to the wide compatibility of materials and integration to microfabrication processes for device development. Many gas plasmas have been used such as air, oxygen, water vapor, ammonia, and argon for modification of polymer surfaces. Plasma modification processes generate new chemical species on polymer surfaces. The new chemical species can arise due to surface reactions with reactive gases or due to physical sputtering (such as with Ar plasma) caused by active gas-phase species. These new surface chemical species can provide an anchor for attaching a series of different molecules that display different properties from the underlying bulk polymer. For example, preferential hydroxylation of poly(methyl methacrylate) or PMMA surfaces by the use of a water-vapor plasma as opposed to an oxygen plasma has been used to activate surfaces toward trichlorosilane modification and subsequent “click” chemistries to form surface scaffolds of desired functionalities. The “click” chemistry method has also been used to modify glass channels for systematic control over electroosmotic flow velocity [5].

Applications

In this section, representative applications utilizing modified surfaces are presented along with short discussions of the underlying physics. Chemically modified surfaces are influenced by changes to surface charge density, as discussed above. Any charged surface in contact with a liquid forms an electric double layer (EDL). A schematic diagram illustrating the classic EDL structure is shown in Fig. 1, depicting the positions to which different characteristic surface charge-related potentials are referenced within the EDL.

For a given substrate, it is useful to quantify the relationship between the surface charge and the surface potential in the presence of EDL. Based on the Gouy-Chapman theory, Grahame’s

equation [6] shown as Eq. 5 addresses this relationship assuming overall electroneutrality (i.e., the sum of surface charge and the total ionic charge, sometimes also referred to as space charge within the device volume, must be zero):

$$\sigma_s = \sqrt{8C_0 \varepsilon \varepsilon_0 k_B T} \operatorname{Sinh}\left(\frac{e\psi_0}{2k_B T}\right) \quad (5)$$

where, σ_s is the surface charge, C_0 is the bulk concentration, ε is the dielectric permittivity, ε_0 is the permittivity of free space, T is temperature, e is the elementary charge, and ψ_0 is the surface potential. Recent results [7] show that actively controlling the surface potential (and consequently the excess surface charge) via embedded electrodes can alter flow phenomenon in nanofluidic channels. By making the electrode potential progressively negative, active flow pumping of aqueous electrolytes was observed, and making the electrode potential progressively positive resulted in reversal of both the bulk flow and electric current [7] for a native wall with a net negative surface charge density.

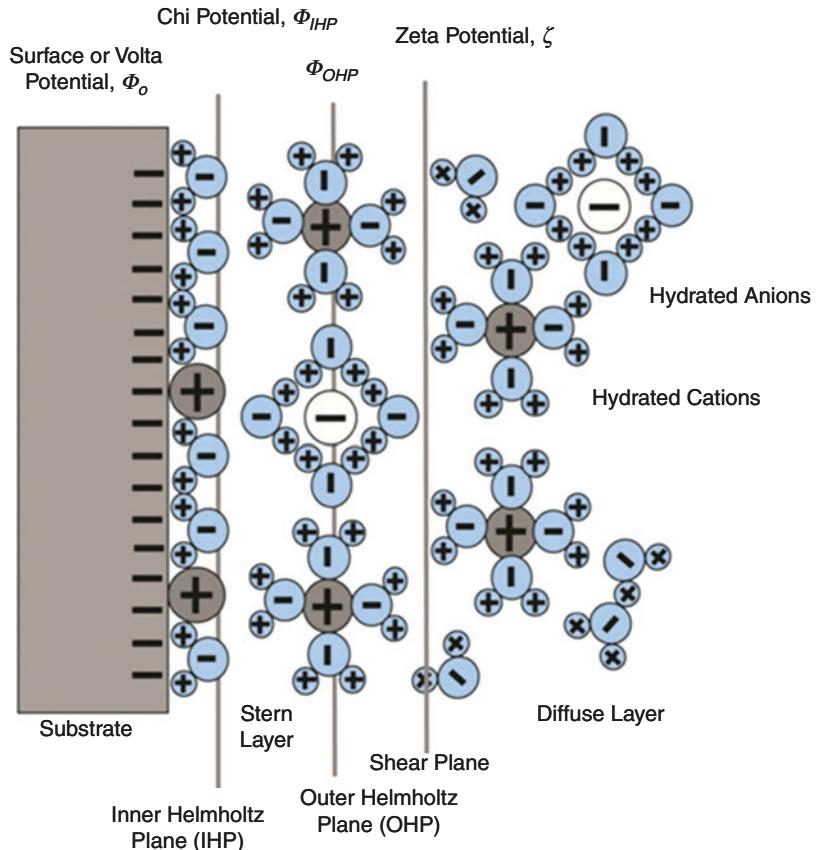
One direct consequence of surface modification is a change in the interaction forces between the surface and the surroundings. Specifically for microfluidics and nanofluidics, changes to surface charge density, surface roughness, and surface energy alter the electrostatic and van der Waals forces that determine flow (ionic and fluid) phenomena. Following the scaling of forces discussed above, electric fields and capillary forces are the two most common tools used to fill microchannels and nanochannels with liquids. For capillary forces, consider the Washburn equation, which provides one approach to quantify the rate of channel filling:

$$\frac{dl_f}{dt} = \frac{a\gamma \cos \theta_c}{4\mu l_f}, \quad (6)$$

where l_f is the fill length, a is the radius (or characteristic length) of the channel, θ_c is the contact angle between the fluid wall and the fluid, γ is the interfacial (or surface) tension, and μ is the fluid viscosity. Equation 6 shows the dependence

Surface-Modified Microfluidics and Nanofluidics,

Fig. 1 Schematic representation of the electric double layer. The surface charge and potentials are depicted based on the theories developed over the past century (Figure from Prakash and Yeom [1])



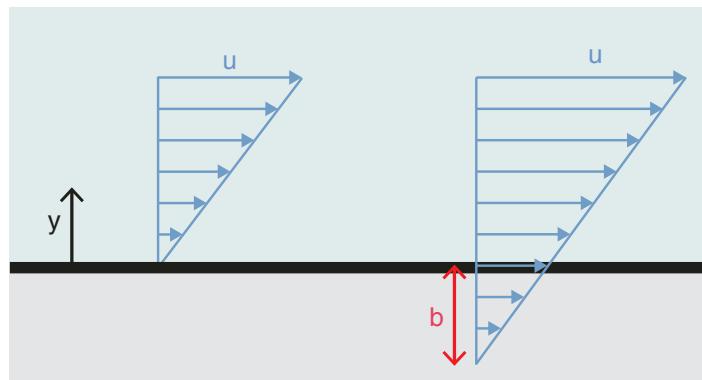
of the capillary (or micro-/nanochannel) filling as a function of surface properties governed by θ_c . Chemical surface modification can easily affect a change to the surface wettability and therefore change θ_c . Therefore, for a given liquid such as water in a micro- or nanochannel, the rate of filling decreases as the surface becomes progressively more hydrophobic.

Given the importance of walls for microfluidics and nanofluidics, a contrast to the boundary conditions with traditional fluid mechanics should be considered. For example, macroscale fluid mechanics assumes a no-slip condition at the wall, which implies that the velocity of the fluid layer in contact with the walls is the same as that of the walls. However, in the early 1800s, Navier postulated that fluid slip maybe possible and suggested the slip boundary condition with u_w , the velocity at the wall (tangential component), can be expressed by the relationship in Eq. 7:

$$u_w = b \frac{\partial u_b}{\partial y} \quad (7)$$

where, b is the slip length, u_b is the bulk velocity, and y is the axis perpendicular to the wall. Figure 2 shows a conceptual schematic for defining the physical effect of slip flow. Conceptually, one can imagine the slip length as the additional distance the wall must be extended to where the fluid velocity would be zero.

Literature has noted that there are three types of slip [8]. First is molecular or intrinsic slip, which occurs when molecular forces (e.g., van der Waals) are balanced by viscous forces. Intrinsic slip is usually observed at very high shear rates (e.g., shear rate of 10^{12} s^{-1} for water). Second is apparent slip in which there is a finite distance between the slip and no-slip planes. Electrokinetically driven microfluidics and nanofluidics follow this type of slip with the slip plane (location where ζ potential is defined in Fig. 1)



Surface-Modified Microfluidics and Nanofluidics,
Fig. 2 Conceptual schematic for the slip boundary condition. The panel on the *left* shows the no-slip boundary condition for a stationary wall, with the wall velocity and

the tangential component of fluid velocity being zero. The *right panel* shows a case for slip flow with a finite slip length (Figure based on [5, 6], and assistance from Mr. M. Hansen in drawing the figure is acknowledged)

being at a finite distance from the physical channel wall. Third is the effective slip which refers to the estimate of either intrinsic or apparent slip by averaging a measurement over the length scale of the experiment. The main factors that affect slip are roughness of the surface and wettability, which presents a measure of the surface energy of the surface. Other possible factors are gases trapped in the fluid, nanobubbles, shear rates of fluid at the surface, and adsorbates in the fluid [8–10], all of which can be influenced by surface modification.

Recently, it was also demonstrated that surface charge can alter the drainage velocity for confined fluids indicating possibly a role for surface charge in mediating slip flow [3]. Additional use of slip phenomena has been in developing drag reduction configurations both for laminar and turbulent flows, self-cleaning super hydrophobic surfaces [11], and enhanced mixing in microfluidic devices [12].

External pressure-driven flow of aqueous solutions in confined micro-/nanochannel configurations (with a finite surface charge) results in the generation of a current and potential and is termed streaming current and streaming potential, respectively [13]. Contrary to the typical no-slip boundary condition, for a given pressure gradient, fluid slip at the walls results in a higher flow rate of the electrolytes. Extrapolating this phenomenon to a nanofluidics battery, a higher flow rate is

commensurate to an increased efficiency of streaming potential generation in the case of slip flows [14].

Over a decade back, research on imaging hydrophobic surfaces immersed in water using atomic force microscopy resulted in the discovery of soft gaseous regions on the surface [15]. Previous research over the last decade has shown that irrespective of the substrate, the height of these bubbles is at least an order of magnitude smaller than its diameter [16]. Among its many applications, the existence of these interfacial gas bubbles helps in increasing the capture efficiency in industrial froth flotation experiments, and subsequent degassing of the liquid resulted in substantial decrease in capture efficiency [17]. Also, formation of nanobubbles on coal surfaces using hydrodynamic cavitation has shown to enhance flotation recovery [18]. Electrochemical formation of nanobubbles at surfaces helps in protein removal and thus prevents surface fouling [19]. The presence of nanobubbles is believed to be one of the reasons for the existence of fluid slip at hydrophobic surfaces as the bubbles offer a low shear surface in comparison to the solid substrate and may therefore be pivotal in drag reduction applications in micro- and nanoscale flows [20]. Moreover, recent advances show that altering the surface hydrophobicity of borosilicate glass via surface modification can significantly alter

nanobubble morphology, density, and the contact line tension [21].

Continuing with a discussion of applications relevant to microfluidics and nanofluidics, valving (or metering of fluids) is one critical operation for sample manipulation for lab-on-chip devices. Valving by surface-mediated flows is a passive method due to a fixed surface state and can be achieved by the tendency of the fluid to favor a channel direction based on the wettability or the surface charge of the channel. Examples include the use of hydrophobic-hydrophilic interfaces to generate boundaries for directing water [22] and use of chemically modified surfaces for exploiting differences in surface charge for selectively driving electrokinetically driven flows in microchannels [23].

Summary

Microfluidic and nanofluidic phenomena are governed by large SA/V ratios. Therefore, interaction between fluids, species, and the channel walls are critical. Toward the development of devices and systems-enabling applications and related fluid control, the ability to systematically modify and control wall properties provides a powerful tool for scientists and engineers.

References

- Prakash, S., Yeom, J.: *Nanofluidics and Microfluidics*. Elsevier (2014)
- Prakash, S., et al.: Nanofluidics: systems and applications. *IEEE Sensors J.* **8**, 441–450 (2008)
- Wu, Y., et al.: Dynamic response of AFM cantilevers to dissimilar functionalized silica surfaces in aqueous electrolyte solutions. *Langmuir* **26**(22), 16963–16972 (2010)
- Prakash, S., Karacor, M.B., Banerjee, S.: Surface modification in microsystems and nanosystems. *Surf. Sci. Rep.* **64**(7), 233–254 (2009)
- Prakash, S., et al.: ‘Click’ modification of silica surfaces and glass microfluidic channels. *Anal. Chem.* **79**(4), 1661–1667 (2007)
- Grahame, D.C.: The electrical double layer and the theory of electrocapillarity. *Chem. Rev.* **41**(3), 441–501 (1947)
- Pinti, M., et al.: Active surface potential control for artificial ion pumps. In: Technical Digest Solid-State Sensors, Actuators, and Microsystems Workshop. Hilton Head Island (2014)
- Lauga, E., Stone, H.A., Brenner, M.P.: Microfluidics: the no-slip boundary condition. In: Tropea, C., Yarin, A., Foss, J.F. (eds.) *Handbook of Experimental Fluid Dynamics*, pp. 1219–1240. Springer, New York (2007)
- Granick, S., Zhu, Y., Lee, H.: Slippery questions about complex fluids flowing past solids. *Nat. Mater.* **2**, 221 (2003)
- Neto, C., et al.: Boundary slip in Newtonian liquids: a review of experimental studies. *Rep. Prog. Phys.* **68**, 2859–2897 (2005)
- Bhushan, B., Jung, Y.C.: Natural and artificial surfaces for superhydrophobicity, self-cleaning, low adhesion, and drag reduction. *Prog. Mater. Sci.* **56**, 1–108 (2011)
- Rothstein, J.P.: Slip on superhydrophobic surfaces. *Annu. Rev. Fluid Mech.* **42**, 89–109 (2010)
- Daiguiji, H., et al.: Electrochemomechanical energy conversion in nanofluidic channels. *Nano Lett.* **4**(12), 2315–2321 (2004)
- Goswami, P., Chakraborty, S.: Energy transfer through streaming effects in time-periodic pressure-driven nanochannel flows with interfacial slip. *Langmuir* **26**(1), 581 (2009)
- Tyrrell, J.W.G., Attard, P.: Images of nanobubbles on hydrophobic surfaces and their interactions. *Phys. Rev. Lett.* **87**(17), 176104 (2001)
- Craig, V.S.J.: Very small bubbles at surfaces—the nanobubble puzzle. *Soft Matter.* **7**(1), 40 (2011)
- Dai, Z., Fornasiero, D., Ralston, J.: Influence of dissolved gas on bubble-particle heterocoagulation. *J. Chem. Soc. Faraday Trans.* **94**(14), 1983–1987 (1998)
- Zhou, Z.A., et al.: On the role of cavitation in particle collection in flotation – a critical review. II. *Miner. Eng.* **22**(5), 419 (2009)
- Wu, Z., et al.: Cleaning using nanobubbles: Defouling by electrochemical generation of bubbles. *J. Colloid Interface Sci.* **328**(1), 10–14 (2008)
- Wang, Y., Bhushan, B., Maali, A.: Atomic force microscopy measurement of boundary slip on hydrophilic, hydrophobic, and superhydrophobic surfaces. *Chin. J. Vac. Sci. Technol. A: Vac. Surf. Films* **27**(4), 754–760 (2009)
- Rangharajan, K.K., Kwak, K.J., Conlisk, A.T., Wu, Y., Prakash, S.: Effect of surface modification on interfacial nanobubble morphology and contact line tension. *Soft Matter*, **11**, 5214–5223 (2015).
- Zhao, B., Moore, J.S., Beebe, D.J.: Surface-directed liquid flow inside microchannels. *Science* **291**(5506), 1023 (2001)
- Prakash, S., Karacor, M.B.: Surface mediated flows in glass nanofluidic channels. In: Solid State Sensors, Actuators, and Microsystems Workshop. Transducer Research Foundation, Hilton Head Island (2010)

Surface-Plasmon-Enhanced Solar Energy Conversion

- Plasmonic Structures for Solar Energy Harvesting

Suspension Interaction with External AC Field Gradient

- AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation

Synchronization

- Nonlinear and Parametric NEMS Resonators

Synthesis of Carbon Nanotubes

Simon J. Henley, José V. Anguita and S. Ravi P. Silva
Nano Electronics Center, Advanced Technology Institute, University of Surrey, Guildford, Surrey, UK

Synonyms

Growth of carbon nanotubes; Growth of CNTs

Definition

Carbon nanotubes (CNTs) are allotropes of graphitic carbon with a cylindrical structure and diameters of <100 nm. A CNT can consist of one or many concentric graphene sheets rolled up as cylinders. CNTs are of interest for a wide

variety of technological applications. In this entry, the various experimental methods to synthesize carbon nanotubes are introduced.

Introduction

Carbon nanotubes (CNTs) were first brought to worldwide attention by Iijima in 1991 [1] after he analyzed, by electron microscopy, the samples produced during an electrical arc discharge between carbon rods held in a helium atmosphere. He observed nanoscale hollow tubes, similar to those seen by Russian researchers in the 1950s [2]. Carbon nanotubes can be visualized as graphene sheets rolled up to form tubes; if one sheet is rolled up, a single-walled carbon nanotube (SWCNT) is formed. The structures formed when two or more concentric tubes are present are termed multiwalled carbon nanotubes (MWCNTs) [3–5]. The high temperatures formed during an electrical arc give an idea as to the extreme conditions that were thought to be required to form these structures. Since then, a wide variety of different growth technologies have been developed covering a broad range of different environmental conditions, from high-temperature laser vaporization down to catalytically enhanced growth in rarefied gas mixtures at much lower temperatures.

Initially, the arc discharge method [2, 6] was mainly employed to produce carbon nanotubes. Later, many more techniques such as laser ablation [7] or chemical vapor deposition (CVD) [8] were developed, researched, and optimized toward the large-scale synthesis of CNTs with repeatable and controllable morphologies. The synthesis of CNTs has now moved from a small-scale research activity to an industrial scale, with production facilities producing hundreds of tons per year. However, CNTs are not only synthesized in laboratories; they can be formed in flames produced by burning organic chemicals such as methane and benzene. They have been found also in soot in the air, likely from industrial or automotive processes, and have been generated by metallurgical process such as those used to make Damascus steel. These naturally occurring CNTs are

typically of low quality though. In this entry the main methods used to synthesize CNTs are introduced.

Synthesis Methods

Arc Discharge

The arc discharge technique (see Fig. 1) generally involves the use of two high-purity graphite electrodes. The electrodes are brought in close proximity, and an electric arc is struck between them by connection to a high-current, low-voltage power supply [3, 6, 9]. Similar electric arcs were used by Roger Bacon in the early 1960s to synthesize large carbon fibers. The synthesis is carried out in a controlled environment at a low pressure (50–700 millibar) of an inert gas, typically helium or argon.

The distance between the electrodes is controlled to keep a large current (50–150 A) flowing between them. The conditions in the region between the electrodes are extreme, and a high-temperature plasma is formed. The temperature is high enough that carbon evaporates (sublimes) from the anode. After the reaction time (typically 30 s to a couple of minutes) has expired and the system has cooled, the products can be collected. The products take different forms depending on the region of the reactor that they are collected from. Carbon soot is found to coat the chamber walls, and a more solid deposit is left on the cathode. The soot contains mainly fullerenes,

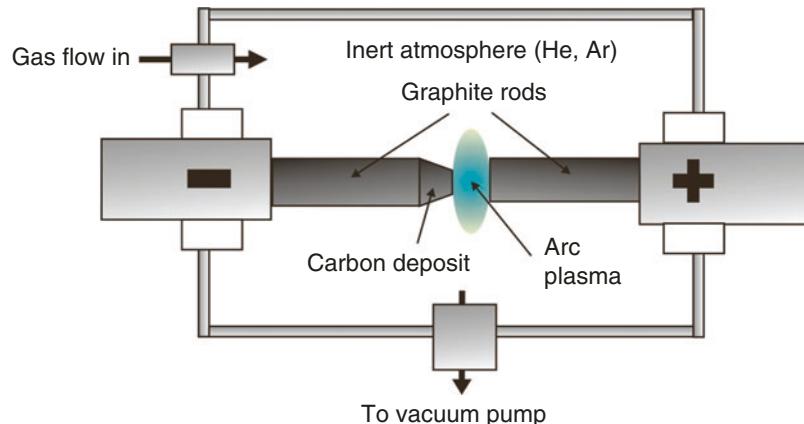
while MWCNTs and graphitic nanoparticles are found in the solid deposit. The yield of the process can be quite high and the CNTs are of high quality, typically, with few defects. The rate of CNT synthesis can be high, producing around 20–100 mg/min [3]. The rate of synthesis has been shown to increase with increasing gas pressure. Although higher currents during the arc would be expected to produce more material, it has been shown that the deposit becomes fused at higher current, actually reducing the effective yield [3, 6]. As no metallic catalyst particles are required for the growth of MWCNTs, arc discharge tubes can be used without acidic treatment to remove the metal.

When metal catalysts (e.g., iron, nickel, or cobalt) are incorporated into the anode and evaporated along with the graphite, the nature of the deposits changes and SWCNTs can be synthesized. The deposit on the electrode is found to contain SWCNTs, MWCNTs, metal-filled MWCNTs, and other graphitic contaminants, while the soot contains MWNTs and SWNTs. In addition, in some situations, a “collar” is observed around the deposit containing mainly SWNTs (80 %) with diameters of around 1 nm. The exact nature of the growth products is determined strongly by the physical conditions during the arc such as the gas pressure, the carbon flux, the arc current, and the composition of the catalyst. Growth models have been proposed by a range of groups in order to try and explain the process [3, 6].

Synthesis of Carbon

Nanotubes,

Fig. 1 Schematic diagram of an arc discharge system to synthesize CNTs



Laser Ablation

The high power densities produced by focusing a short-duration laser pulse onto a target lead to a very rapid heating of the material at the focal point. Subsequent thermal and other direct-sputtering processes ablate material from the target, and the interaction between the laser pulse and the ablated material forms a highly energetic, nonequilibrium, plasma plume [10]. The presence of ambient gas can have a dramatic effect on the expansion dynamics of these laser-ablated plasmas. At higher pressures plume confinement and plume heating occur. The high collision rates enhance conversion of the plume stream velocity into thermal energy, producing a confined, high-temperature plasma, with short mean-free paths for species. In this regime, the self-assembly of nanoclusters in the gas phase can occur. By ablating a target containing a mixture of the growth material and a suitable catalyst and by controlling the gas chemistry, nanomaterials such as CNTs can be self-assembled in the gas phase.

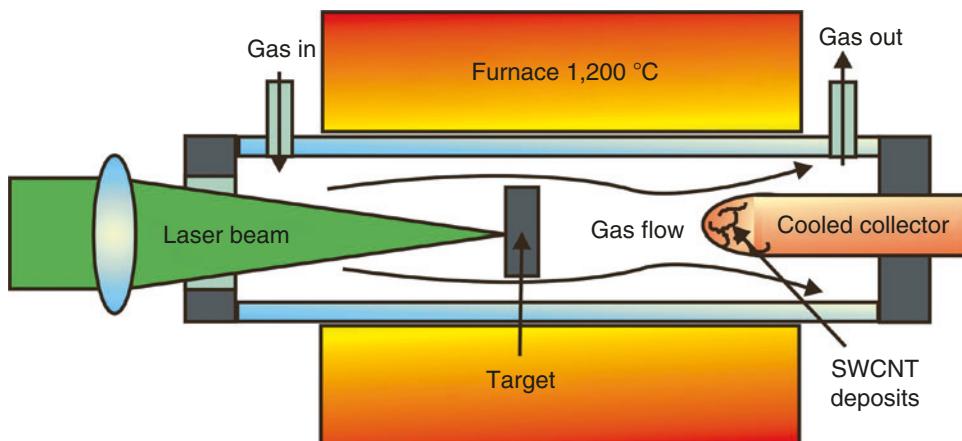
In the laser ablation CNT synthesis process (see Fig. 2), a pulsed laser vaporizes a graphite target held at a high temperature (typically 1,000–1,200 °C) inside a tube, within a furnace. An inert gas flows through the tube driving the ablation products toward one end. SWCNTs that grow in the gas phase are deposited in the cooler parts of the reactor as the vaporized

carbon condenses. A water-cooled collector may be included in the system to collect the CNTs [3, 5, 9].

The laser ablation synthesis method was demonstrated by Smalley et al. [7, 11] where the growth of MWCNTs was demonstrated. Later this method was developed further to produce SWCNTs by ablating graphite targets mixed with metallic elements such as cobalt and nickel [3, 11–13]. The laser ablation method with a catalyst present has a yield around 70 % and produces primarily single-walled carbon nanotubes with a diameter determined by the temperature of the furnace. The quality of the SWCNTs is very high, although there is typically a significant quantity of catalyst nanoparticles present in the product, which may need to be removed by acid treatment producing additional defects.

Chemical Vapor Deposition

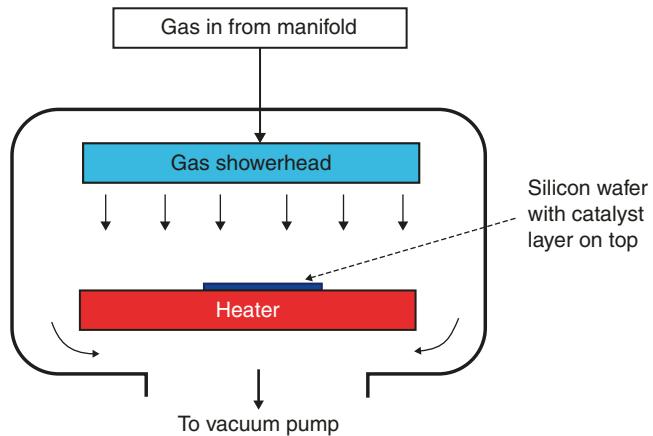
Chemical vapor deposition (CVD) is a vacuum deposition technique that allows the deposition of materials uniformly over large surface areas (typically flat surfaces) with high levels of purity. For this reason, CVD equipment (and its variants) is commonly found in semiconductor-manufacturing foundries and is used for depositing thin layers of materials on silicon wafers. For CNT growth, a catalyst is often used, in the form of a thin metal layer on top of the substrate, in



Synthesis of Carbon Nanotubes, Fig. 2 Schematic diagram of an arc discharge system to synthesize CNTs

Synthesis of Carbon Nanotubes,

Fig. 3 Schematic of a CVD system



order to lower the temperature that is required for CNT growth [3, 8, 9]. This method is sometimes referred to as catalytic CVD (CCVD) in the literature due to the involvement of catalytic activity during the CVD growth process [3, 8]. In this method, the substrate plus catalyst is mounted on a heated stage, inside a vacuum chamber. A schematic of the CVD arrangement is shown in Fig. 3.

The temperature of the substrate is raised, typically between 400 and 1,000 °C in a low pressure of hydrogen gas. At elevated temperatures, the catalyst thin film breaks up spontaneously and forms nanoparticles on the surface of the substrate. A carbon-containing gas, typically a hydrocarbon such as acetylene or methane, is then bled into the chamber, together with a carrier gas, typically hydrogen. Hydrocarbon concentrations of typically between 1 % and 10 % are used. Upon contact with the catalyst nanoparticles, the hydrocarbon gas decomposes and releases carbon to the individual nanoparticles. The nanoparticles then extrude this carbon in the form of concentric graphitic sheets (the nanotube). A dynamic equilibrium is set up at the nanoparticles between the hydrocarbon decomposition rate and the growth speed of the CNTs, which ensures the continual growth of the CNTs. The exact growth mechanism and events that take place at the nanoparticles during growth are still the subject of debate [8].

Although commercial CVD equipment has been available for more than half a century for the deposition of traditional semiconductors and

dielectrics, the development of systems solely for CNT growth is only recent, and commercial systems (e.g., Surrey NanoSystems' NanoGrowth system) have only been released to the market within the last 5 years. This is due to the more recent discovery of the techniques that allow for growth processes that are compatible with the already existing semiconductor infrastructure. In particular, techniques for reducing the growth temperatures of CNTs have been sought, as silicon processing only allows maximum operating temperatures of up to around 450 °C before degradation of the silicon diffusion layers starts to take place [14, 15].

Water-Assisted CVD

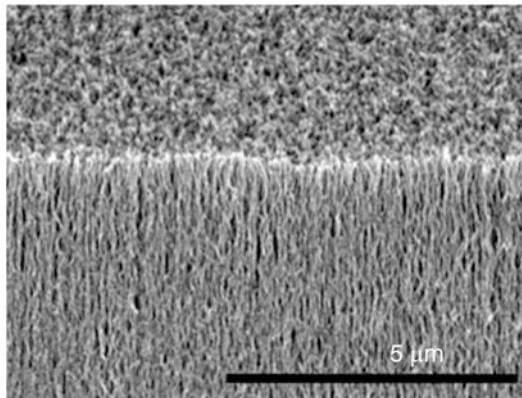
Another area that is still subject of debate is the mechanism by which CNT growth termination occurs when using CCVD techniques. It is noted that CNT growth stops abruptly sometime after growth initiation, typically after around 5–10 min. It is believed that this takes place due to the buildup of amorphous carbon at the surface of the catalyst particles, which effectively “poisons” the catalytic properties of the particles, inhibiting the dynamic equilibrium required for CNT growth. This theory was enlightened when samples of CNTs where growth had terminated were cleaned using an oxygen plasma to remove this amorphous carbon. It was noted that CNT growth could resume after performing this step.

Despite this success, it was clear that introducing oxygen gas as a cleaning agent during CNT

growth (which involves hydrogen and hydrocarbon gas mix) would be unsafe. For this reason water vapor is currently used for this purpose, which although less effective than oxygen at removing amorphous carbon is compatible with the CNT growth gas mix. It is now well accepted that introducing a small amount of water vapor to the hydrogen and hydrocarbon gas mix reduces the growth termination effect (see Fig. 4). This allows growing CNTs in lengths up to several millimeters, reaching the centimeter scale, in a single run [16].

Plasma-Enhanced CVD

A further variation of the CVD process is plasma-enhanced CVD (PECVD) [17]. In this case, a



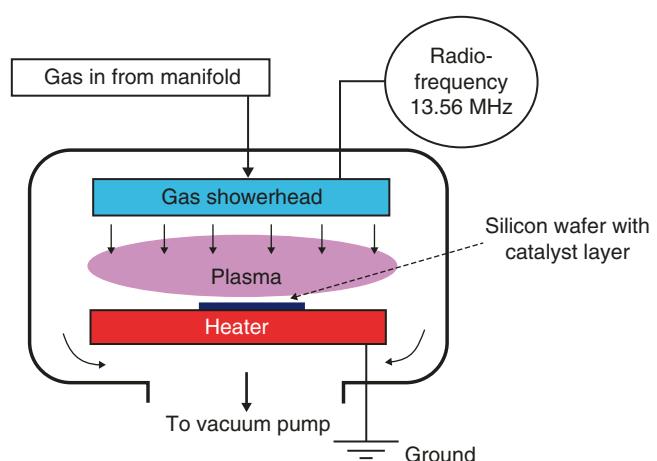
Synthesis of Carbon Nanotubes, Fig. 4 SEM image of long CNTs grown by water-assisted CVD

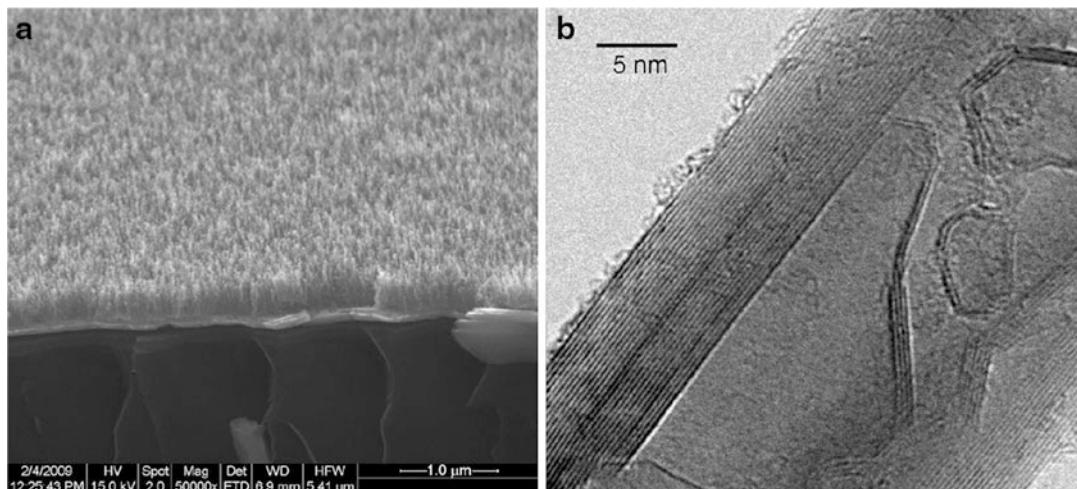
Synthesis of Carbon Nanotubes,

Fig. 5 Simplified schematic of a PECVD system

plasma is struck between the gas showerhead and the substrate electrode (with the heater), by means of the application of a radio frequency to the showerhead electrode. This arrangement is depicted in Fig. 5. In this case, a fraction of the hydrocarbon gas decomposes in the plasma phase to produce smaller molecules and energetic species which provide additional energy at the catalytic nanoparticles, allowing a further reduction of the temperature of the substrate. Additionally, this plasma provides a strong electric field between the surface of the substrate and the plasma sheath (a region of the plasma that is located only a few millimeters above the substrate electrode). This electric field provides a vertically upward pull force on the nanoparticles, which results in a high degree of CNT verticality, as depicted in Fig. 6a. This allows for the formation of vertical CNT “forests” even when the CNT density is low. Although (PE)CVD can produce large quantities of CNTs, the quality of the material is lower than that produced by arc discharge and laser ablation, with many structural defects present in the tubes (see Fig. 6b).

As for the case of CVD, PECVD process equipment is also common in silicon-manufacturing foundries and has been commercially available for many decades. However, specific equipment dedicated to CNT growth has only been available in recent years. This is partly because of specialized techniques that are involved in striking a stable plasma at the higher





Synthesis of Carbon Nanotubes, Fig. 6 (a) SEM image of CNTs grown using the PECVD process, showing a high degree of verticality. (b) High-resolution TEM

image of a CVD MWCNT showing internal structural defects and surface contaminants

gas pressures required for CNT growth, typically 5 Torr and above, up to several hundred Torr. These pressures are much higher than those typically associated with the common PECVD processes, of a few hundred millitorr.

Fluidized Bed CVD

The CVD process is a promising route for the bulk production of CNTs; however, producing truly industrial-scale volumes of material is challenging. Fluidized bed CVD (FBCVD) reactors are becoming one of the most widely targeted methods to achieve these scales. However, scale-up of the reactor is still a significant challenge, and this field has become an active area of research [18, 19]. In the FBCVD process, the CVD reaction occurs within a fluidized bed of catalyst particles. The setup consists of a reaction tube inside a cylindrical furnace. The contiguous mixing that occurs in the catalyst beds allows efficient utilization of all of the catalyst particles, which are all exposed to the feed gases. One such catalytic process, CoMoCAT®, is briefly introduced, but a detailed discussion of the mechanisms can be found in a recent review article [19].

The CoMoCAT® technique uses active cobalt stabilized in a nonmetallic state by interaction

with molybdenum oxide (MoO_3) in a fluidized bed. When the mixture is exposed to carbon monoxide, the Co-Mo dual oxide is carburized, producing molybdenum carbide and small Co nanoparticles which are of very uniform sizes and well dispersed, allowing the growth of SWCNTs with small diameters and a high selectivity for certain chiral indices.

Floating Catalyst CVD

Similar to the fluidized bed CVD, a high degree of catalyst utilization can be achieved if the nanoparticles are kept suspended in the gas phase. If a narrow size distribution of small catalyst nanoparticles can be generated, growth of SWCNTs can be achieved. This idea leads to a simple method for production whereby the catalyst is introduced into a CVD reactor either by (1) a syringe process using the catalyst dissolved in a carbon source, (2) by sublimation of the catalyst at elevated temperatures, or (3) using a gaseous catalyst source, e.g., $\text{Fe}(\text{CO})_5$. The catalyst and the carbon source are directly reacted in the gas phase. One such process that has attracted much attention is the HiPco technique, as it has been scaled to produce industrial quantities of SWCNTs. We discuss this process in more detail here.

The high-pressure carbon monoxide disproportionation process (HiPco) is a method for the production of SWCNTs using a continuous flow of high-pressure carbon monoxide as the carbon source and an iron carbonate $\text{Fe}(\text{CO})_5$ as the source of catalyst nanoparticles [20]. This process was developed at Rice University by the group of the late R.E. Smalley. SWCNTs are produced by flowing CO, mixed with a small amount of $\text{Fe}(\text{CO})_5$ through a heated reactor. The HiPco process produces SWCNTs with diameters of approximately 1.1 nm. The yield achieved is approximately 70 % SWCNTs with 97 % purity at rates of up to 450 mg/h [20].

Summary of Growth Mechanisms

In this entry the main routes to synthesize CNTs have been detailed. A summary of the important points is shown in Table 1.

Purification

The technical challenges in the production of usable CNTs often do not end when the material

is first grown. After growth the yield and purity of the product is often not high enough for direct utilization; so, purification to remove unwanted contaminants such as metallic catalyst particles and non-CNT carbonaceous materials is required. The techniques employed typically are strong gaseous oxidation and chemical treatments such as acid refluxing, which will both have an effect (often detrimental) on the structure of the tubes [3]. Following the demonstration that carbon nanotubes could be attacked by oxidizing gases, Ebbesen et al. realized that the more defective carbon nanoparticle contaminants, such as amorphous carbon, would be oxidized more readily than the more perfect nanotubes. They found that a significant purification of CNTs could be achieved this way, but with a significant loss in yield [6]. For the chemical purification routes, a typical treatment method would involve (a) dispersal in organic solvent and filtration to remove large particulates, (b) treatment with concentrated acids to remove fullerenes and catalyst particles, (c) centrifugal separation, and (d) microfiltration. Such a chemical treatment can cause damage to the surface layer of MWCNTs and possible total destruction of SWCNTs, if too extreme.

Synthesis of Carbon Nanotubes, Table 1 Summary of the three main growth methods

Method	Arc discharge	Laser ablation	CVD
<i>CNT type</i>	<i>SWCNTs</i> : short nanotubes with diameters of around a nanometer	<i>SWCNTs</i> : formation of bundles of long (many microns) nanotubes, with diameters of a few nm	<i>SWCNTs</i> : growth of isolated tubes with diameters in the range of 1–4 nm
	<i>MWCNTs</i> : short nanotubes with inner diameters of a few nm and outer diameter of around 10 nm	<i>MWCNTs</i> : not typically grown by this method	<i>MWCNTs</i> : growth of long (many microns) nanotubes with diameters from a few nm up to hundreds of nm (carbon fibers)
<i>Yield</i>	30–90 %	Up to 70 %	Up to 100 %
<i>Pros</i>	Relatively simple equipment required and produces large quantities of SWCNTs or MWCNTs. CNTs produced are of a good quality, with few defects. Yield is reasonable	SWCNTs produced by this method have the highest quality of all the methods and diameter control is good. Yield is good	Can be grown over large areas and in large quantities (good for industrial scale-up). Diameters can be controlled relatively easily by size of catalyst. Low-temperature growth possible. Purity can be high
<i>Cons</i>	Size control is difficult and significant purification is often required	Expensive experimental setup (high-peak power lasers). Purification is often required	Nanotubes are highly defective typically

Cross-References

- Carbon Nanotube-Metal Contact
- Carbon Nanotubes
- Carbon Nanotubes for Chip Interconnections
- Functionalization of Carbon Nanotubes

References

1. Iijima, S.: Helical microtubules of graphitic carbon. *Nature* **354**, 56 (1991)
2. Radushkevich, L.V., Lukyanovich, V.M.: Zurn. Fisic. Chim. **26**, 88 (1952)
3. Harris, P.F.: Carbon Nanotube Science: Synthesis, Properties and Applications, 2nd edn. Cambridge University Press, Cambridge (2009)
4. Ebbesen, T.W. (ed.): Carbon Nanotubes, Preparation and Properties. CRC Press, Boca Raton (1996)
5. Saito, R., Dresselhaus, G., Dresselhaus, M.S. (eds.): Physical Properties of Carbon Nanotubes. World Scientific, Singapore (1998)
6. Ebbesen, T.W.: Carbon nanotubes. *Annu. Rev. Mater. Sci.* **24**, 235 (1994)
7. Guo, T., Nikolaev, P., Rinzler, A.G., Tománek, D., Colbert, D.T., Smalley, R.E.: Self-assembly of tubular fullerenes. *Phys. Chem.* **99**, 10694–10697 (1995)
8. Kumar, M., Ando, Y.: Chemical vapor deposition of carbon nanotubes: a review on growth mechanism and mass production. *J. Nanosci. Nanotechnol.* **10**, 3739–3758 (2010)
9. Dresselhaus, M.S., Dresselhaus, G., Avouris, P. (eds.): Carbon Nanotubes: Synthesis, Structure, Properties, and Applications. Topics in Applied Physics, vol. 80. Springer, Berlin (2001)
10. Ashfold, M.N.R., Claeysens, F., Fuge, G.M., Henley, S.J.: Pulsed laser ablation and deposition of thin films. *Chem. Soc. Rev.* **33**, 23–31 (2004)
11. Yakobson, B.I., Smalley, R.E.: Fullerene nanotubes: C-1,000,000 and beyond. *Am. Sci.* **85**, 324 (1997)
12. Guo, T., Nikolaev, P., Thess, A., Colbert, D.T., Smalley, R.E.: Catalytic growth of single-walled nanotubes by laser vaporization. *Chem. Phys. Lett.* **243**, 49–54 (1995)
13. Thess, A., Lee, R., Nikolaev, P., Dai, H., Petit, P., Robert, J., Xu, C., Lee, Y.H., Kim, S.G., Rinzler, A.G., Colbert, D.T., Scuseria, G.E., Tománek, D., Fischer, J.E., Smalley, R.E.: Crystalline ropes of metallic carbon nanotubes. *Science* **273**, 483 (1996)
14. Boskovic, B.O., Stolojan, V., Khan, R.U.A., Haq, S., Silva, S.R.P.: Large area synthesis of carbon nanofibres at room temperature. *Nat. Mater.* **1**, 165–168 (2002)
15. Chen, G., Jensen, B., Stolojan, V., Silva, S.R.P.: Growth of carbon nanotubes at temperatures compatible with integrated circuit technologies. *Carbon* **49**, 280–285 (2011)
16. Amama, P.B., Pint, C.L., McJilton, L., Kim, S.M., Stach, E.A., Murray, P.T., Hauge, R.H., Maruyama, B.: Role of water in super growth of single-walled carbon nanotube carpets. *Nano Lett.* **9**, 44–49 (2009)
17. Meyyappan, M., Delzeit, L., Cassell, A., Hash, D.: Carbon nanotube growth by PECVD: a review. *Plasma Sources Sci. Technol.* **12**, 205–216 (2003)
18. See, C.H., Harris, A.T.: A review of carbon nanotube synthesis via fluidized-bed chemical vapor deposition. *Ind. Eng. Chem. Res.* **46**, 997–1012 (2007)
19. MacKenzie, K.J., Dunens, O.M., Harris, A.T.: An updated review of synthesis parameters and growth mechanisms for carbon nanotubes in fluidized beds. *Ind. Eng. Chem. Res.* **49**, 5323–5338 (2010)
20. Nikolaev, P.: Gas-phase production of single-walled carbon nanotubes from carbon monoxide: a review of the hipco process. *J. Nanosci. Nanotechnol.* **4**, 307 (2004)

Synthesis of Functional Materials for Bone Regeneration

Oscar Castaño

Biomaterials for Regenerative Therapies, Institute for Bioengineering of Catalonia (IBEC), Barcelona, Spain

Synonyms

Bioactive materials; Osteogenic materials; Osteoinduction materials

Definitions

3D scaffold processing methods: These are techniques to implement 3D porous scaffolds in order to obtain proper mechanical properties, as well as suitable porosity, to be able to colonize the whole structure with cells. Examples include foaming, freeze-drying, sintering, salt leaching, rapid prototyping, electrospraying, etc., [1].

Angiogenesis: The creation of new blood vessels from preexisting ones. It is commonly accepted that cells cannot be further than 150–200 µm from the closest blood capillary to avoid inducing necrosis by hypoxia.

Bioactivity: The ability to trigger a specific biological response. In bone tissue regeneration, a material is usually considered to have bioactive ability when it is possible to carefully precipitate bone-like hydroxyapatite onto its surface after immersion in an aqueous serum-like solution. However, a serious controversy exists among the scientific community because this process leads to many false positive and false negative results [2].

Bioceramic: An inorganic ceramic material, a crystal, a glass or a glass-ceramic, that is intended to be in contact with living tissues. In bone regeneration, bioceramics are used to regenerate and repair living tissues damaged by disease or trauma. They play a key role in specific clinical applications in orthopedics and dentistry or drug delivery [3]. Typical bioceramics are hydroxyapatite, polymorphs α and β of tricalcium phosphate ($\text{Ca}_3(\text{PO}_4)_2$), and silicon-based (bioglass) and titanium-based bioactive glasses.

Biocompatibility: The ability of a material to be parenterally implanted with a suitable host response in developing a specific function.

Biodegradability: The ability to be naturally dissolved and transformed into new compounds by hydrolysis, oxidation, enzymes, bacteria, or osteoclasts [4].

Biodegradable biomaterial: A material that, once parenterally applied, can be naturally dissolved, or enzymatically degraded, and its byproducts released into the bloodstream. Ideally, a biodegradable material would also be eliminated by the human body by natural pathways and without toxicity. Examples are calcium phosphates, polylactic acid, and polyglycolic acid.

Bone: A strong and tough connective tissue that supports and protects the rest of the internal organs, allows the body to move, offers an ideal environment for blood cell formation, and acts as a store for salts (especially calcium and phosphates), among other functions [1]. The bone, as a hybrid natural construct, presents a complex and highly hierarchical organized structure [5]. Morphologically, it can be subdivided into two distinct types according to porosity and unit microstructure: the cortical bone (also known as compact or dense bone), which represents around 80 % of the total bone mass, and the trabecular bone (also

called cancellous or spongy bone) with a high surface area and porosity enough for the bone marrow and blood vessels [6].

Bone is basically the combination of two main phases: An inorganic one, mainly formed by carbonated hydroxyapatite, which represents around 60 % of the total mass, and an organic constituent, mainly formed by collagen type I organized in fibril bundles and other minor proteoglycans, non-collagenous macromolecules. Finally, cells complete the organic composition, forming a highly hierarchical organized 3D network together.

Bone extracellular matrix: A complex collagen-based fibrous structure that includes proteins, proteoglycans, and other signals and structure molecules. It contains the necessary signals to guide bone cells in the restoration of structure and function of damaged or dysfunctional bone tissue by surface mineralization [7].

Chemotaxis: The ability to recruit a particular type of cells by chemical signals. In bone regeneration, special emphasis is given to the promotion of the transferring and participation of MSCs to the damaged site [8].

Confluence: Part of the culture surface that is covered by a monolayer of cells.

Extracellular matrix (ECM): In mammals, a collagen-based structure expressed by cells as one of their main functions that acts as guiding template support and biochemical signaling platform for the development of other functions such as cell adhesion, proliferation, cell-cell signaling, or differentiation. Composition is very variable depending on the mammal species and tissues, but in general, it is basically formed by collagen fibrils that provide other proteins (fibronectin, elastin, laminin, etc.), growth factors, glycosaminoglycans, and other physical and chemical signals.

Extracellular matrix signaling: A combination of chemical and physical stimuli related to the extracellular matrix, providing a three-dimensional extracellular microenvironment. Physical stimuli include topography; mechanical stimuli include stiffness or external loads; and chemical stimuli include molecules, usually proteins or steroids that can be plasma soluble or

non-soluble, usually related to the extracellular matrix by some type of bond.

Growth factor: Signaling protein or hormone that is able to regulate cell response by matching with the proper receptor on the cell membrane.

Instructive material: An engineered device providing the required signals on the surface or in the body of a material to be able to interact and control cell tissue response in order to promote bone healing [7]. Signals can be chemical (chemical gradients, ion and biomolecule release, functional groups on the surface, etc.), physical (roughness, anisotropy, topography gradients, engineered nanofeatures), or mechanical (stiffness, strain, external active load, mechanical gradients, etc.).

Osteoblasts: Mononuclear cells whose origin is differentiation from mesenchymal progenitor cells. They are responsible for the nucleation, growth, and mineralization of a new bone by the control of the local extracellular matrix ionic calcium and phosphate concentrations. They modulate hydroxyapatite formation using a previously synthesized extracellular matrix, which acts as a template and is mainly formed by collagen I and several types of glycoproteins and chemical and physical signals. At the end of the process, osteoblasts auto-mineralize their membrane and are trapped alive in the neophyte bone, becoming an osteocyte [9].

Osteoclasts: Multinucleated cells whose main function is the degradation and resorption of the bone through local acidification by an enzymatic reaction.

Osteoconduction: The ability of a surface to act as template and allow the growth of bone tissue by facilitating attachment, proliferation, migration, and osteoblast differentiation to form new tissue [10].

Osteocyte: Mineralized and differentiated osteoblasts embedded within the bone matrix. They are smaller than osteoblasts and have lost many of their cytoplasmic organelles [6].

Osteogenesis: The ability to create new bone tissue.

Osteoinduction: The ideal osteoinductive material should be able to induce ectopic bone synthesis when implanted into extra-skeletal body places; in other words, it should be osteogenic in soft tissues [11].

Scaffold: A 3D structure that acts as a guidance template for bone regeneration and generally involves macroporosity ($\varnothing > 50 \mu\text{m}$) to allow cell migration and vascular colonization, micro- and nanoporosity for permeability and, if required, better degradation.

Surface modification: Engineering processes that involve determined change in the surface of a material in order to properly interact with the host tissue cells.

Vasculogenesis: The process that involves the creation of new blood vessels by the new formation of endothelial cells from mesoderm cell precursors [12].

Introduction

Bone tissue regeneration has been one of the driving forces in the development of biomaterials and surface functionalization for the production of instructive materials. The knowledge acquired has formed a basis for assays in other tissues.

Functional materials for bone tissue date from the 1980s, emerging particularly in second-generation biomaterials [13]. Functional materials for bone tissue substitution or regeneration require several features related to mechanical and bioactive characteristics whose main goal is to guide tissue growth. This has enabled the development of a third generation of biomaterials and has allowed the field to advance toward short- to mid-term expectations for smart biomaterials and instructive scaffolds to guide the growth of bone tissue, as well as other human tissues [14, 15].

The tendency nowadays in functional biomaterials is to mimic what is known as extracellular matrix (ECM), including its dynamic cell control characteristics and hierarchical organization [14]. Osteoid, for example, is the result of a process involving osteoblast mineralization of an ECM. Depending on the degree of mineralization, osteoid Young's modulus can range from $\sim 27 \text{ kPa}$ to $\sim 1 \text{ GPa}$, this value being characteristic of mature bone [16]. Today, a controversy exists about the mechanical rigidity of functional biomaterials for bone regeneration: should the initial graft be as rigid as bone, or

should the final result be as stiff as bone, while the biomaterial mimics less rigid, osteoid characteristics?

What seems to be clear is that the design of the material should be tailored to the final application. There are many different bones within the body, and their architecture and composition, even when similar, depend on their specific role and function. This is why there is no universal bone substitute but rather particular designs for the development of specific functions.

Bone

The bone is a hard connective tissue that acts as a protector of other internal organs, contributes to movement support, stores calcium and phosphates, and provides a porous environment for blood cell dynamics, among other specific functions [17]. Bones can be morphologically divided into flat, long, irregular, and short bones [6]. In addition, based on their microstructure, they can also be divided into two main families: cortical or compact bone (dense, low porosity, and void spaces, representing about 80 % of the total bone mass normally located in the outer shell of bone structures for mechanical support) and trabecular, cancellous, or spongy bone [6]. Compact and cancellous bones have a similar hierarchical composition and organized structure [5]. It is mainly formed from a mineral phase (carbonated hydroxyapatite, among other minor compounds) and an organic phase (collagen, among others) both contributing to its exceptional mechanical properties that combine rigidity and load-bearing strength with elasticity and flexibility [5]. Parameters such as composition, crystallinity, constituent morphology, 3D conformation, and distribution vary depending on the bone and its specific function within the body.

3D Scaffolds to Control Cell Fate

It is well known that cells need a material template with the correct 3D chemical and physical guidance to develop their functions, and 3D scaffolds

are the basis for tissue engineering approaches [18]. Such scaffolds' architecture and, in particular, their surface features are directly responsible for governing cell fate by affecting cell functions. Therefore, the design of 3D architectures at the nanoscale which are able to specifically interact with cells' receptors is of utmost importance. The strict biocompatibility requirements can rule out about 90 % of available materials. Once the biological mechanisms involved in the regenerative process are identified, the generated knowledge can lead to the design of a biomaterial specifically tailored toward the control of stem cell fate, thus producing the desired phenotype [19].

Scaffold Design

The design of scaffolds requires:

- The correct mechanical properties to maintain the scaffold's structure during its life.
- A shape that can easily adapt to the damaged cavity.
- Optimal biocompatibility.
- Degradation at the same rate as new bone tissue growth.
- Sterilizability, without undesirable degradation or cross-linking.
- Enough porosity to allow cell and blood vessel colonization, the supply of nutrients and oxygen, and waste elimination. Porosity is measured at three levels: macro-, meso-, and microporosity. Macroporosity refers to pores with a width larger than 50 nm, mesoporosity to those between 2 and 50 nm, and microporosity to those with a width smaller than 2 nm [20].
- Optimal mass and fluid transport through the scaffold and good interconnections between pores.
- Be sterilizable.

Macroporosity is a crucial factor that directly controls cell behavior, bone growth and vascularization, as well as the mechanical properties of scaffolds. The optimal pore size required for bone regeneration depends on the cell lineage, differing efficiency depending on the upper limit of functionality [21]. The porosity of the material is a

Synthesis of Functional Materials for Bone Regeneration, Table 1 List of different sequences with bioactivity effect on cells

Peptide sequence	Derived from	Proven function
RGD	Multiple proteins	Relevant to bone engineering; RGD enhances cell adhesion and differentiation into bone, cartilage, neural, and endothelial tissue
FHRIIKA	Heparin-binding domain	Increases osteoblast adhesion and mineralization
KRSR	Heparin-binding	Increases osteoblast adhesion and mineralization
YIGSR	Laminin	Increases human foreskin fibroblast adhesion
IKVAV	Laminin	Increases neurite extension
REDV	Fibronectin	Increases endothelial cell adhesion
KHIFSDDSSE	Neural cell adhesion molecules	Increases astrocyte adhesion
VPGIG	Elastin	Increases stiffness of synthetic matrices
SVSVGGMKPSPPR	Phage display	High selectivity toward hydroxyapatite and tooth enamel
VTKHLNQISQSY	Phage display	High selectivity toward hydroxyapatite and bone-like mineral (BLM)
CRKRLDRNC	Phage display	Interacts with IL-4 receptor on endothelial cells, macrophages, and smooth muscle cells
THRTSTLDYFVI	Phage display	High selectivity toward polypyrrole that increases adhesion of PC12 cells
STFTKSP	Phage display	Homes to primitive hematopoietic progenitor cells in bone marrow
WYRGRL	Phage display	Binds to collagen II α 1
ASSLINA	Phage display	Binds to both skeletal and cardiac tissue
CAGALCY	Phage display	Targets brain tissue

Adapted from [28] with kind permission from Springer Science and Business Media

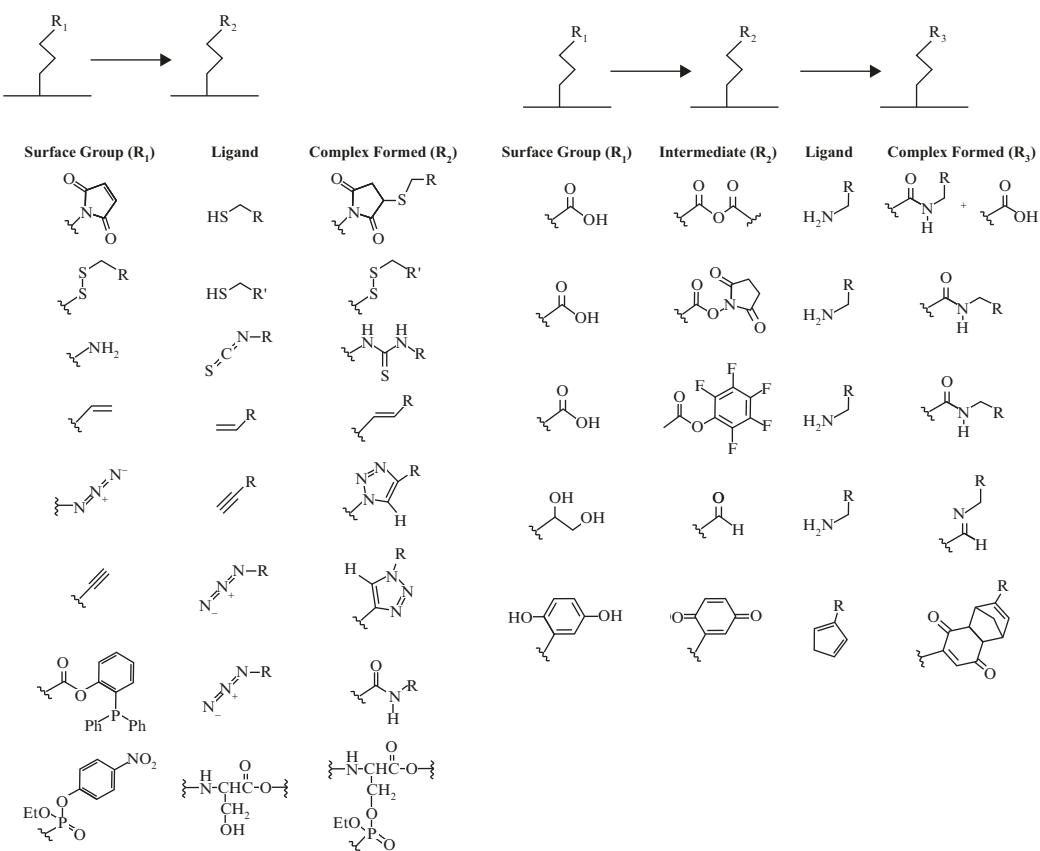
delicate balance between biological behavior and mechanical properties, as the higher the void volume, the better the vascularization and the osteointegration, but the worse the scaffold's resistance to mechanical failure [22]. The interconnectivity of pores is essential for bone regeneration as it enables the infiltration of bone, the development of an efficient network of blood vessels, and the promotion of cell-cell interactions [23]. Specific surfaces also help to have a better understanding of the kinetic degradation of the construct, minimizing usually undesired diffusion control, and this is closely related to the degree of porosity.

On the other hand, the surface morphology also affects cellular response in the form of small pores, patterns, or roughness. Adhesion is one of the favored processes and depends on the design of surface [24]. The strength of the anchorage of cells further determines aspects such as migration, proliferation, or differentiation. Adhesion is one aspect that creates more controversy among

scientists, as weak adhesion leads to cell apoptosis, and very strong adhesion leads to early differentiation before the cells proliferate. As a result, it seems that a medium adhesion strength is ideal, allowing proliferation until the point when a cell will differentiate, for example, and/or a dense confluence (another crucial factor to describe cell behavior is the cell-to-cell communication). The surface can be modulated by texturing and patterning using different methods [25]: blasting, electropolishing, chemical treatments [26], lithography, plasma treatment, and focused ion beam, among others. This can positively influence cell response in terms of adhesion detachment, spreading, migration, and proliferation, which are better at guiding cell behavior.

The nature of the functional groups present on the surface can also be relevant in modulating cell response. Aspects such as hydrophilicity, surface charge, and the surface energy are directly related to the chemical composition of the interface. On the other hand, bioactive biomolecules can also be

Synthesis of Functional Materials for Bone Regeneration, Table 2 Common strategies in surface covalent functionalization: (left) direct link; (right) link through intermediates



Adapted with permission from [29]. Copyright 2015 American Chemical Society

added on the surface in order to induce a biological response in cells. This mainly relies on the use of ECM proteins (collagen, fibronectin, laminin, etc.), peptide sequences that are known to affect specific functions (RGD, YIGSR, REDV, IKVAV, KRSR, etc.), growth factors (BMP, BMP2, VEGF, TNF- α , etc.), drugs (antibiotics, analgesics, antitumor agents, biphosphonates, etc.), dyes (rhodamine, fluorescein, etc.), or bioactive nanoparticles (hydroxyapatite, bioglasses, ormoglasses, etc.) [27]. The strength of the surface binding would depend on the chosen strategy; the biomolecule can remain on the surface once the cell attaches or, alternatively, be released to the media to be detected by the cell (Table 1).

The binding of the biomolecule to the surface of the scaffold can be weak or strong, by physiadsorption or chemiadsorption. In the second case, there are many options depending on the type of surface being functionalized: EDC/NHS-based chemistry, click chemistry, self-assembled monolayers, electrostatic union, etc. [29, 30] (Table 2).

Many processing methods have been developed to shape biomaterials depending on the application. No particular one is perfect, but all provide special feature that can be brought into play for specific roles. Foaming allows the possibility of obtaining a very high porosity, but pore window junctions can be an unsolved problem for

cell migration; sintering is an option only available for crystalline ceramics; salt leaching is an economical way of obtaining high porosity and good interconnectivity, but with poor mechanical properties and thick, slowly degrading walls; rapid prototyping also results in high porosity combined with an optimized mechanical resistance, but also has the problem of thick walls in the range of tens of micrometers; electrospinning is the cheapest method to mimic ECM and one of the methods preferred by cells, but with a poor macroporosity that makes colonization difficult.

Cross-References

- Angiogenesis
- Bioderived Smart Materials
- Bioinspired Synthesis of Nanomaterials
- Biomimetics
- Bone Remodeling
- Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells
- Electrospinning
- Nanoengineered Hydrogels for Cell Engineering
- Nanopatterned Surfaces for Exploring and Regulating Cell Behavior
- Nanostructured Functionalized Surfaces
- Nanostructures for Surface Functionalization and Surface Properties
- Nanotechnology
- Surface Modeling of Ceramic Biomaterials

References

1. Sachot, N., Engel, E., Castaño, O.: Hybrid organic–inorganic scaffolding biomaterials for regenerative therapies. *Curr. Org. Chem.* **18**(18), 2299–2314 (2014)
2. Bohner, M., Lemaitre, J.: Can bioactivity be tested in vitro with SBF solution? *Biomaterials* **30**, 2175–2179 (2009). doi:10.1016/j.biomaterials.2009.01.008
3. Vallet-Regi, M.: Bio-Ceramics with Clinical Applications. Southern Gate, Chichester, West Sussex, UK (2014)
4. Bandyopadhyay, A., Bernard, S., Xue, W., Böse, S.: Calcium phosphate-based resorbable ceramics: influence of MgO, ZnO, and SiO₂ dopants. *J. Am. Ceram. Soc.* **89**, 2675–2688 (2006). doi:10.1111/j.1551-2916.2006.01207.x
5. Clarke, B.: Normal bone anatomy and physiology. *Clin. J. Am. Soc. Nephrol.* **3**, 131–139 (2008). doi:10.2215/CJN.04151206
6. Pérez-Amidio, S., Engel, E.: Bone biology and regeneration. In: *Bio-Ceramics with Clinical Applications*, pp. 315–42. Southern Gate, Chichester, West Sussex, UK (2014). doi:10.1002/9781118406748.ch11
7. Lutolf, M.P., Hubbell, J.A.: Synthetic biomaterials as instructive extracellular microenvironments for morphogenesis in tissue engineering. *Nat. Biotechnol.* **23**, 47–55 (2005)
8. Sordi, V.: Mesenchymal stem cell homing capacity. *Transplantation* **87**, S42–S45 (2009). doi:10.1097/TP.0b013e3181a28533. [pii] r00007890-200905151-00004
9. Bose, S., Tarafder, S.: Calcium phosphate ceramic systems in growth factor and drug delivery for bone tissue engineering: a review. *Acta Biomater.* **8**, 1401–1421 (2012). doi:10.1016/j.actbio.2011.11.017
10. Vacanti, C.A., Pietrzak, W.S.: *Musculoskeletal Tissue Regeneration: Biological Materials and Methods*. Humana Press, Totowa, New York (2008)
11. Miron, R.J., Zhang, Y.F.: Osteoinduction: a review of old concepts with new standards. *J. Dent. Res.* **91**, 736–744 (2012). doi:10.1177/0022034511435260
12. Risau, W., Flamme, I.: Vasculogenesis. *Annu. Rev. Cell Dev. Biol.* **11**, 73–91 (1995). doi:10.1146/annurev.cb.11.110195.000445
13. Dorozhkin, S.V.: Calcium orthophosphates and human beings: a historical perspective from the 1770s until 1940. *Biomatter* **2**, 53–70 (2012)
14. Holzapfel, B.M., Reichert, J.C., Schantz, J.-T., Gbureck, U., Rackwitz, L., Nöth, U., et al.: How smart do biomaterials need to be? A translational science and clinical point of view. *Adv. Drug Deliv. Rev.* **65**, 581–603 (2013). doi:10.1016/j.addr.2012.07.009
15. Castano, O., Sachot, N., Xuriguera, E., Engel, E., Planell, J.A., Park, J.-H., et al.: Angiogenesis in bone regeneration: tailored calcium release in hybrid fibrous scaffolds. *ACS Appl. Mater. Interfaces* (2014). doi:10.1021/am500885v
16. Engler, A.J., Sen, S., Sweeney, H.L., Discher, D.E.: Matrix elasticity directs stem cell lineage specification. *Cell* **126**, 677–689 (2006)
17. Bandyopadhyay-ghosh, S.: Bone as a collagen-hydroxyapatite composite and its repair. *Trends Biomater. Artif. Organs* **22**, 116–124 (2008)
18. Stapor, P.C., Azimi, M.S., Ahsan, T., Murfee, W.L.: An angiogenesis model for investigating multi-cellular interactions across intact microvascular networks.

- Am. J. Physiol. Heart Circ. Physiol. **304**, H235–H245 (2013). doi:10.1152/ajpheart.00552.2012
19. Provenzano, P.P., Keely, P.J.: Mechanical signaling through the cytoskeleton regulates cell proliferation by coordinated focal adhesion and Rho GTPase signaling. *J. Cell Sci.* **124**, 1195–1205 (2011). doi:10.1242/jcs.067009
20. Rouquerol, J., Avnir, D., Fairbridge, C.W., Everett, D.H., Haynes, J.M., Pernicone, N., et al.: Recommendations for the characterization of porous solids. *Pure Appl. Chem.* **66**, 1739–1758 (1994)
21. O'Brien, F.J., Harley, B.A., Yannas, I.V., Gibson, L.J.: The effect of pore size on cell adhesion in collagen-GAG scaffolds. *Biomaterials* **26**, 433–441 (2005). doi:10.1016/j.biomaterials.2004.02.052
22. Hollister, S.J.: Porous scaffold design for tissue engineering. *Nat. Mater.* **4**, 518–524 (2005). doi:10.1038/nmat1421
23. Valerio, P., Guimarães, M.H.R., Pereira, M.M., Leite, M.F., Goes, A.M.: Primary osteoblast cell response to sol-gel derived bioactive glass foams. *J. Mater. Sci. Mater. Med.* **16**, 851–856 (2005). doi:10.1007/s10856-005-3582-5
24. Dalby, M.J., Gadegaard, N., Tare, R., Andar, A., Riehle, M.O., Herzyk, P., et al.: The control of human mesenchymal cell differentiation using nanoscale symmetry and disorder. *Nat. Mater.* **6**, 997–1003 (2007). doi:10.1038/nmat2013
25. Gadegaard, N., Dalby, M.J., Martines, E., Seunarine, K., Riehle, M.O., Curtis, A.S.G., et al.: Nano patterned surfaces for biomaterial applications. *Adv. Sci. Technol.* **53**, 107–115 (2006)
26. Montanaro, L., Arciola, C.R., Campoccia, D., Cervellati, M.: In vitro effects on MG63 osteoblast-like cells following contact with two roughness-differing fluorohydroxyapatite-coated titanium alloys. *Biomaterials* **23**, 3651–3659 (2002)
27. Sachot, N., Castaño, O., Mateos-timoneda, M.A., Engel, E., Planell, J.A.: Hierarchically engineered fibrous scaffolds for bone regeneration hierarchically engineered fibrous scaffolds for bone regeneration. *J. R. Soc. Interface* **10**, 1–5 (2013)
28. Segvich, S., Kohn, D.: Phage display as a strategy for designing organic/inorganic biomaterials. In: Puleo, D. A., Bizios, R. (eds.) *Biological Interactions on Materials Surfaces SE – 6*, pp. 115–32. Springer, New York (2009). doi: 10.1007/978-0-387-98161-1_6
29. Love, J.C., Estroff, L.A., Kriebel, J.K., Nuzzo, R.G., Whitesides, G.M.: Self-assembled monolayers of thiolates on metals as a form of nanotechnology. *Chem. Rev.* **105**, 1103–1170 (2005). doi:10.1021/cr0300789
30. Conde, J., Dias, J.T., Grazú, V., Moros, M., Baptista, P.V., de la Fuente, J.M.: Revisiting 30 years of biofunctionalization and surface chemistry of inorganic nanoparticles for nanomedicine. *Front. Chem.* **2**, 48 (2014). doi:10.3389/fchem.2014.00048

Synthesis of Gold Nanoparticles

Munish Chanana¹, Cintia Mateo¹, Verónica Salgueirino² and Miguel A. Correa-Duarte¹

¹Departamento de Química Física, Universidade de Vigo, Vigo, Spain

²Departamento de Física Aplicada, Universidade de Vigo, Vigo, Spain

Synonyms

Production of gold nanoparticles

Definition

Description of different synthetic approaches for the fabrication of gold nanoparticles.

Overview

Since the beginning of recorded history, gold has always held the majestic throne among all the noble metals [1, 2]. It has been known by artisans since the Chalcolithic (Copper) Age and has become a highly coveted metal for coinage, jewelry, and other arts since that time. The first extraction of gold has been dated back to the fifth millennium B.C. in Bulgaria, but “soluble” gold (colloidal gold) first appeared around the fifth century B.C. in Egypt and China [1, 2]. Colloidal gold was used to make ruby glass and to color ceramics. Perhaps the most famous examples are the Lycurgus Cup (manufactured around fifth/fourth century B.C., exhibited in British Museum) and the pigment “Purple of Cassius” (invented by Andreas Cassius, seventeenth century) [1, 2]. But scientific research on gold sol started with Michael Faraday. In 1857, Faraday reported the formation of deep red solutions of colloidal gold by the reduction of an aqueous solution of chloroaurate (HAuCl_4) using phosphorus in CS_2 (a two-phase system) [1, 2].

The most relevant properties of gold colloids are based on the presence of a strong absorption band in the visible-NIR, which is the origin of the observed brilliant red/purple colors of certain gold nanoparticles in solution. This absorption band results from the collective oscillation of the conduction band electrons in resonance with the frequency of the incident electromagnetic field and is known as surface plasmon resonance (SPR) absorption. The SPR frequency and thus the color of the gold NP mainly depends on the particle size, shape, the nature of the surrounding medium, and the interparticle distance. The influence of shape and interparticle distance is in general even greater than that of size [1, 2]. While a single absorption band is present for spherically symmetric gold particles, multiple absorption bands correlated with their various axes appear for nonspherical ones. Such structures can support both propagating and localized surface plasmons. For instance, gold nanorods possess two different resonance modes due to electron oscillation across and along the long axis of the nanorod and are commonly labeled the transverse and longitudinal modes, respectively, the latter of which is extremely sensitive to the aspect ratio of the rod [3]. Gold nanotriangles, for instance, exhibit three different resonances in the UV–Vis absorption spectra, one SPR absorption out of plane and two in-plane SPR absorptions at longer wavelengths [4, 5]. In the case of branched, platonic, or platelet nanostructures, the plasmon resonance shifts to longer wavelengths with respect to the common resonance of spherical nanoparticles at 520 nm [4–6]. Special attention requires the case of gold nanoshells which, usually due to their synthetic approach, are composed of a dielectric core coated with a thin metallic layer. Although these structures have similar properties to spherical gold nanoparticles in terms that they only present a single SPR absorption, they offer, however, the ability of tuning the SPR over a full range of wavelengths from the visible to the infrared region [4–6].

Since particle shape has a tremendous effect on the physical, chemical, optical, electronic, and catalytic properties of nanoparticles, gold nanoparticles have been synthesized in various

shapes, including rods, cubes, plates, polyhedrons, and wires following different and ingenious techniques [1–8]. In the literature, there are multiple excellent reviews [1–6] and books [7, 8] regarding the synthesis of gold nanoparticles of different sizes and shapes, where their different physical properties (e.g., optical, electronic, or mechanical) are also discussed. Therefore, in this chapter, we provide a comprehensive overview on the major wet-chemistry synthetic approaches for the preparation of gold nanoparticles attending to the particle morphology. Hence, the main criterion on which this chapter is structured is the shape of the nanoparticles, predominantly according to the isotropy of the particles.

Isotropic Gold Nanoparticles

Spheres

Sphere is the most thermodynamically favored shape for all kinds of colloidal particles synthesized in bulk solutions. Spherical gold nanoparticles can be synthesized in solution in various sizes, ranging from 1 nm up to several hundreds of nanometers, and with different capping agents. The capping agent mainly dictates the chemical and physicochemical behavior of the particles, and this in turn determines the employment of the particles [1, 2].

One of the most popular synthetic methods for the preparation of gold nanospheres is based on the reduction of HAuCl₄ by citrate in water, which was first described in 1951 by Turkevich [1, 2]. In this method, citrate serves as both a reducing agent and an anionic stabilizer. It yields 15 nm nanoparticles with a fairly narrow size distribution. The Turkevich method has been modified by a number of groups to produce spherical gold nanoparticles with diameters ranging from 15 to 150 nm by either controlling the ratio of citrate to HAuCl₄ or employing a γ -radiation method. However, nanoparticles larger than 20 nm synthesized by this method usually lack isotropy and size uniformity [1, 2].

Other reduction methods have also been developed to achieve a better control over the size and

monodispersity, including the “Schmid method” published in 1981 and the “Brust-Schiffrin method” reported in 1994 [1, 2]. The Schmid’s cluster $[Au_{55}(PPh_3)_{12}Cl_6]$ remained unique for a long period of time with its narrow dispersity (1.4 ± 0.4 nm) for the study of a quantum-dot nanomaterial, despite its arduous synthesis. The synthesis requires rigorously anaerobic conditions and diborane gas as a reducing agent. As a result, phosphine-stabilized gold nanoparticles have lost favor since the development of a more convenient, scalable synthesis of thiol-stabilized air-stable gold nanoparticles of reduced polydispersity and controlled size by the method of Brust et al. [1, 2].

In the Brust-Schiffrin method, the synthesis is carried out in a two-phase system and thiol ligands that strongly bind to gold $AuCl_4^-$ are transferred to toluene using tetraoctylammonium bromide as the phase-transfer reagent and reduced by $NaBH_4$ in the presence of dodecanethiol. The organic phase changes color from orange to deep brown within a few seconds upon addition of $NaBH_4$. The particle size ranges between 1.5 and 5.2 nm. Additionally, these gold NPs can be repeatedly isolated and dispersed in common organic solvents without any aggregation or decomposition. Using the principle of this synthesis method, modified procedures have been developed to synthesize gold nanoparticles in the size range of 4–10 nm capped with thiol-functionalized molecules (small molecules or polymers) in one phase system using methanol as the solvent [1, 2].

The use of microemulsions, copolymer micelles, and inverse micelles is a significant research field for the synthesis of stabilized gold NPs. Typically, these syntheses involve a two-phase system with a surfactant that causes the formation of the microemulsion or the micelle maintaining a suitable microenvironment of controlled dimensions. The surfactants create small pockets of a water phase in an organic solvent or vice versa where the surfactant faces the aqueous phase with its polar group and the tail faces the organic phase. Varying the ratio of the two solvents affects the dimensions of the micelles, which allows for tuning the size of the resulting nanoparticles with good monodispersity [1, 2].

All of the methods described above lead to uniform nanoparticles, however, only in the size range of below 20 nm. For the synthesis of larger spherical NPs (Fig. 1), the seeded-growth procedure is the most popular technique that has been intensively studied in recent years [1, 2]. A strong chemical reducing agent (usually sodium borohydride) is used to generate small, generally spherical, nanoparticles (commonly denoted as *seeds*), which are then added to a growth solution composed of surfactants and Au metal ions to induce particle growth. The growth solution employs a weaker reducing agent (often ascorbic acid) to reduce the metal salt to an intermediate state so that only catalyzed reduction on the nanoparticle surface is allowed, avoiding secondary nucleation. During the growth process, usually a small amount of undesired nonspherical particles is also generated, which can be removed through a CTAB-assisted shape-selective separation method [1, 2, 4]. The gold nanospheres obtained after such purification can be further grown by means of the same procedure described here up to a diameter of 200 nm or more. The final size of the nanospheres synthesized by this method depends strongly on the size of the employed seeds and the ratio between the seeds and the gold salt. To achieve predetermined final sizes of the nanospheres, the concentrations can be adjusted using the following equation [9]:

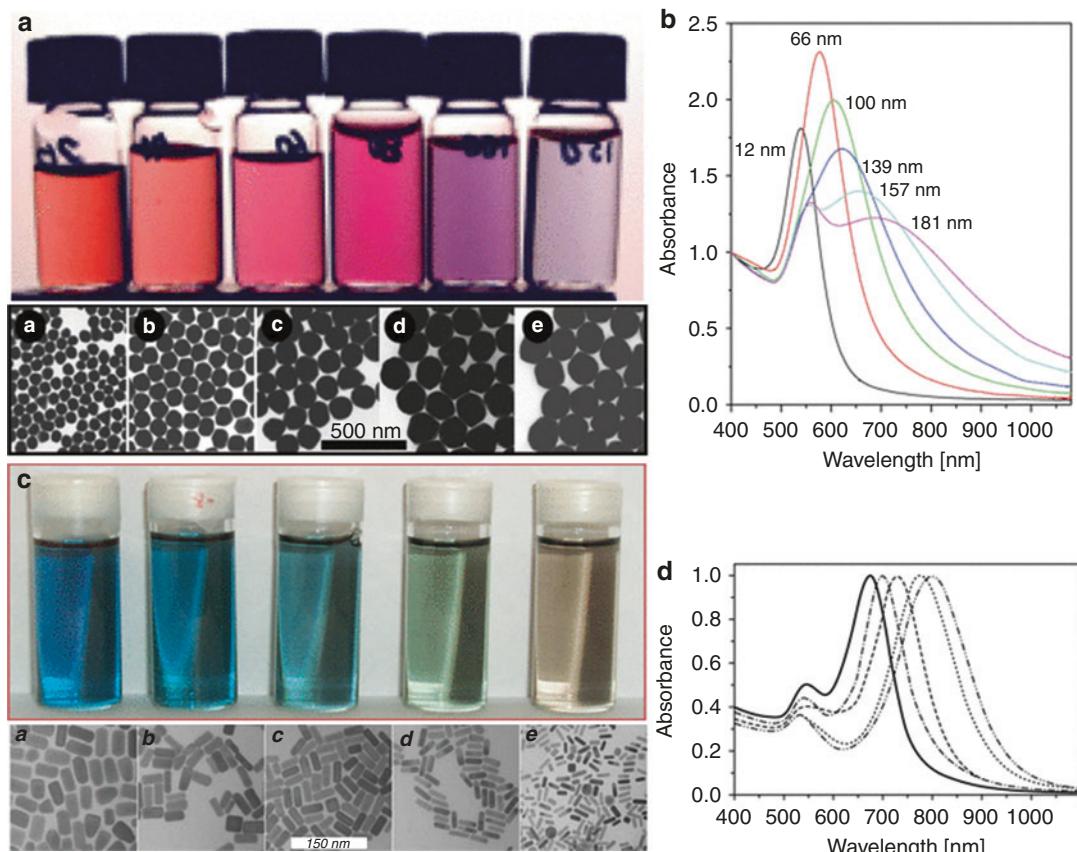
$$R_{\text{final}} = R_{\text{seeds}} \left(\frac{[Au^{3+}]_{\text{final}} + [Au_{\text{seeds}}]_{\text{final}}}{[Au_{\text{seeds}}]_{\text{final}}} \right)^{\frac{1}{3}}$$

where $[Au_{\text{seeds}}]_{\text{final}}$ is the value to be calculated:

$$[Au_{\text{seeds}}]_{\text{final}} = - \frac{[Au^{3+}]_{\text{final}}}{1 - (R_{\text{final}}/R_{\text{seeds}})^3}$$

Shells

Nanoshells have similar properties to spherical gold nanoparticles in the sense that they also exhibit a single surface plasmon resonance (SPR) absorption. However, the SPR absorption band of nanoshells can be tuned across the visible and infrared region over a range of wavelengths spanning hundreds of nanometers, far exceeding



Synthesis of Gold Nanoparticles, Fig. 1 (a) Representative TEM micrographs of Au spheres (*bottom*) obtained after subsequent growth steps with their respective color solution (*top*). Average diameters are 66 (*a*), 100 (*b*), 139 (*c*), 157 (*d*), and 181 (*e*) nm. (b) UV-Vis spectra of Au spheres with various average diameters (*right*) (Reprinted with permission from Ref. [24]. Copyright 2006 American Chemical Society.) (c) Photographs (*top*)

and the respective TEM micrographs (*bottom*) of gold nanorod solutions with average aspect ratios of 1.94 (*a*), 2.35 (*b*), 2.48 (*c*), 3.08 (*d*), and 3.21 (*e*). (d) UV-Vis absorption spectra of different gold nanorod samples stabilized with CTAB with increasing aspect ratio from *left to right* (Reprinted with permission from Ref. [3, 25]. Copyright 2004 Elsevier B.V.)

the spectral range of spherical particles. This has a huge advantage for SERS-based applications, because the plasmon resonance can be tuned to the excitation of common laser radiation sources optimizing the electromagnetic enhancement mechanism [5, 10, 11]. Apart from their optical properties, these nanostructures also present inert biological activities which make them appropriate for biomedical applications, such as targeted drug delivery, photothermal therapeutic applications, and molecular imaging, as contrast agents for optical coherence tomography (OCT).

Nanoshells are usually composed of dielectric cores (e.g., silica particles) coated with a single or

multiple thin metal layers [10]. The growth of metal nanoshells on core particles combines techniques of molecular self-assembly with the reduction chemistry of metal colloid synthesis. This approach is general and can potentially be adapted to a variety of core and shell materials. The common synthesis route of gold nanoshells involves the synthesis of the dielectric core material, typically silica or polystyrene, and their surface functionalization with terminal amine groups to facilitate attachment of small colloidal gold, which are usually formed in a separate process and act as nucleation sites for the subsequent gold salt reduction and shell growth. The absorbed

seed colloids increase in size as reduction ensues, followed by the coalescence of gold particles on the surface, until finally the apparent formation of a polycrystalline continuous metallic nanoshell occurs [10].

Although this common two-step process of seeding, i.e., formation of small colloidal gold followed by attachment to the core, is effective, it is also laborious and costly, especially in terms of time. Another alternative and also very effective approach is the deposition-precipitation (DP) method, which employs the *in situ* formation of gold seeds directly on the support. The functionalized core particles are placed in contact with a basic solution of gold (III) chloride, and upon heating, oxidic precursor particles ($\text{Au}[\text{OH}]_3$) are formed on the support. These precursor seed particles then act as nucleation sites for the formation of gold nanoshells. The size and density of the seeds on the substrate is easily controlled by the concentration of gold (III) chloride, support surface functionalization, reaction temperature and time, and the pH of reaction. This DP method is relatively easy to handle and is currently used for producing commercial gold catalysts in a highly controlled manner [11].

Anisotropic Gold Nanoparticles

Since the last century, colloid chemists have gained excellent control over the synthesis of spherical gold particles in terms of size and uniformity. However, synthetically controlling particle shape, particularly in solutions, has been a bigger challenge and has met somewhat limited success. Nevertheless, especially in the last 10 years, various methods have been developed to synthesize a variety of anisotropic gold nanoparticles with somewhat controllable shape and morphology. Anisotropic gold nanoparticles have been synthesized in symmetric shapes, such as cylindrical, cubic, polyhedral, and plate-like shapes, but more interestingly, there are methods that produce predominately asymmetric branched morphologies such as nanostars. In the following, the most representative synthesis techniques for various types of anisotropic gold nanoparticles

will be described in ascending order of shape complexity [3–6].

Nanorods

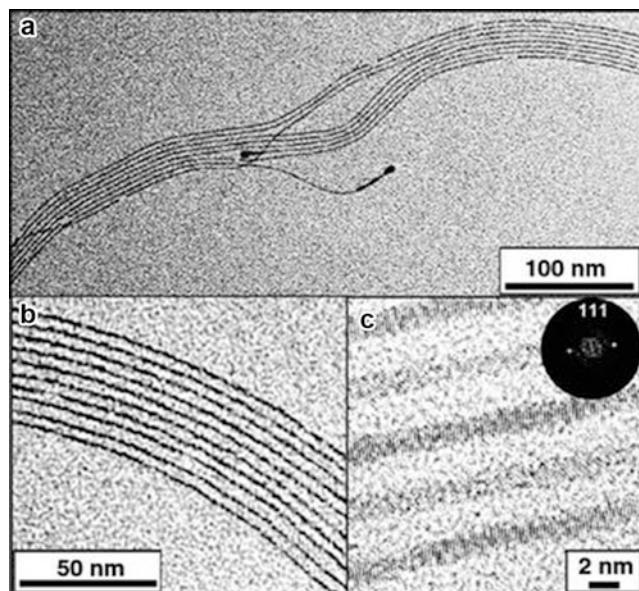
In the field of anisotropic nanoparticles, the synthesis of gold nanorods is perhaps the most established protocol in terms of the degree of control of the size, shape, and monodispersity [3, 4]. There are three main methods used to produce gold rods through wet chemistry, namely, the template method, electrochemical method, and seeded-growth method [3]. The template method for the synthesis of gold nanorods is based on the electrochemical deposition of Au within the pores of nanoporous polycarbonate or alumina template membranes. For this, Ag or Cu is first sputtered onto the alumina template membrane to provide a conductive film for electrodeposition, i.e., the growth of Au nanorods. Subsequently, both the template membrane and the copper or silver film are selectively dissolved in the presence of a polymeric stabilizer such as poly(vinylpyrrolidone) (PVP), which allows for dispersing the rods either in water or in organic solvents by means of sonication or agitation. The dimensions of these synthesized nanorods coincide with the pore diameter of the template membranes which can be tuned by controlling the pore sizes [3].

The electrochemical synthesis of gold nanorods is conducted within a simple two-electrode-type electrochemical cell, where a gold metal plate is used as a sacrificial anode while the cathode is a platinum plate with similar dimensions. Both electrodes are immersed in an electrolytic solution containing a mixture of cationic surfactants, hexadecyltrimethylammonium bromide (CTAB), tetradodecylammonium bromide (TDDAB), acetone, and cyclohexane. The electrolytic cell containing the mixture is then placed inside an ultrasonic bath at 36 °C and controlled-current electrolysis performed with a typical current of 3 mA for a typical electrolysis time of 30 min. The particle sizes can be tuned either via electrolysis time or by addition of silver ions (silver plate) [3].

The seeded-growth method is perhaps the most facile and most applied method to synthesize gold nanorods. During the early attempts of

Synthesis of Gold Nanoparticles,

Fig. 2 TEM images of Au nanowires at different magnifications (**a, b**) and an HRTEM image (**c**) showing the single-crystalline structure (Reprinted with permission from Ref. [13]. Copyright 2008 American Chemical Society)



seed-mediated approaches of growing presynthesized spherical gold nanoparticles in solution, formation of a distinct population (5–10%) of colloidal gold rods was observed. In the beginning of this century, it was found that by controlling the growth conditions in aqueous surfactant media, it was possible to synthesize gold nanorods with tunable aspect ratio (Fig. 1). Moreover, addition of AgNO_3 influences not only the yield and aspect ratio control of the gold nanorods but also the mechanism for gold nanorod formation and correspondingly its crystal structure morphology and optical properties. A fine-tuning of the aspect ratio of the nanorods can be achieved by simply varying the pH in silver-free syntheses or by adjusting one of the parameters in the silver-mediated procedures, namely, the concentration of silver ions, gold ions, seeds, or reducing agent (ascorbic acid), while keeping the other parameters constant [3, 4].

Nanowires

In comparison to nanorods, nanowires exhibit much larger aspect ratios, and hence larger anisotropy, which makes them ideal candidates for many applications, such as biosensors, logic circuits, field-effect transistors, and nonvolatile

memory elements [4, 5]. Being conductive materials, metal nanowires have driven a multitude of theoretical studies on their conductance behavior in the fabrication of nanoelectronics. Inspired from the synthesis of gold nanorods, various methods have been developed to synthesize gold nanowires. These include the pore-template method, block-polymer or DNA template methods, patterned lithography, and chemical reduction methods in bulk solutions. Seed-mediated growth in aqueous CTAB solutions [12] and seedless growth in oleylamine solutions of organic solvents are prime examples of the chemical synthesis of gold nanowires in solutions (Fig. 2).

Depending on the synthesis method and the growth mechanism, gold nanowires can feature different dimensions (diameter, length, and aspect ratio), crystallinity (monocrystalline or polycrystalline or twinned), and shape (straight or twisted). Inspired by the seeded-growth method, Kim et al. reported an acidic solution route to synthesize gold nanowires at room temperature with diameters tunable between 16 and 66 nm and lengths up to 10 μm with an aspect ratio larger than 200 [12]. On the other hand, the method described by Giersig and coworkers, [13] ensures

the synthesis of thin gold nanowires with diameters of ~ 1.6 nm and lengths ranging from 10 nm to ≥ 3.5 μm . Control over Au wire lengths is realized by tuning the oleylamine/HAuCl₄ volume ratio, the reaction time, and the addition of a second solvent.

Platonic Nanoparticles (Polyhedrons)

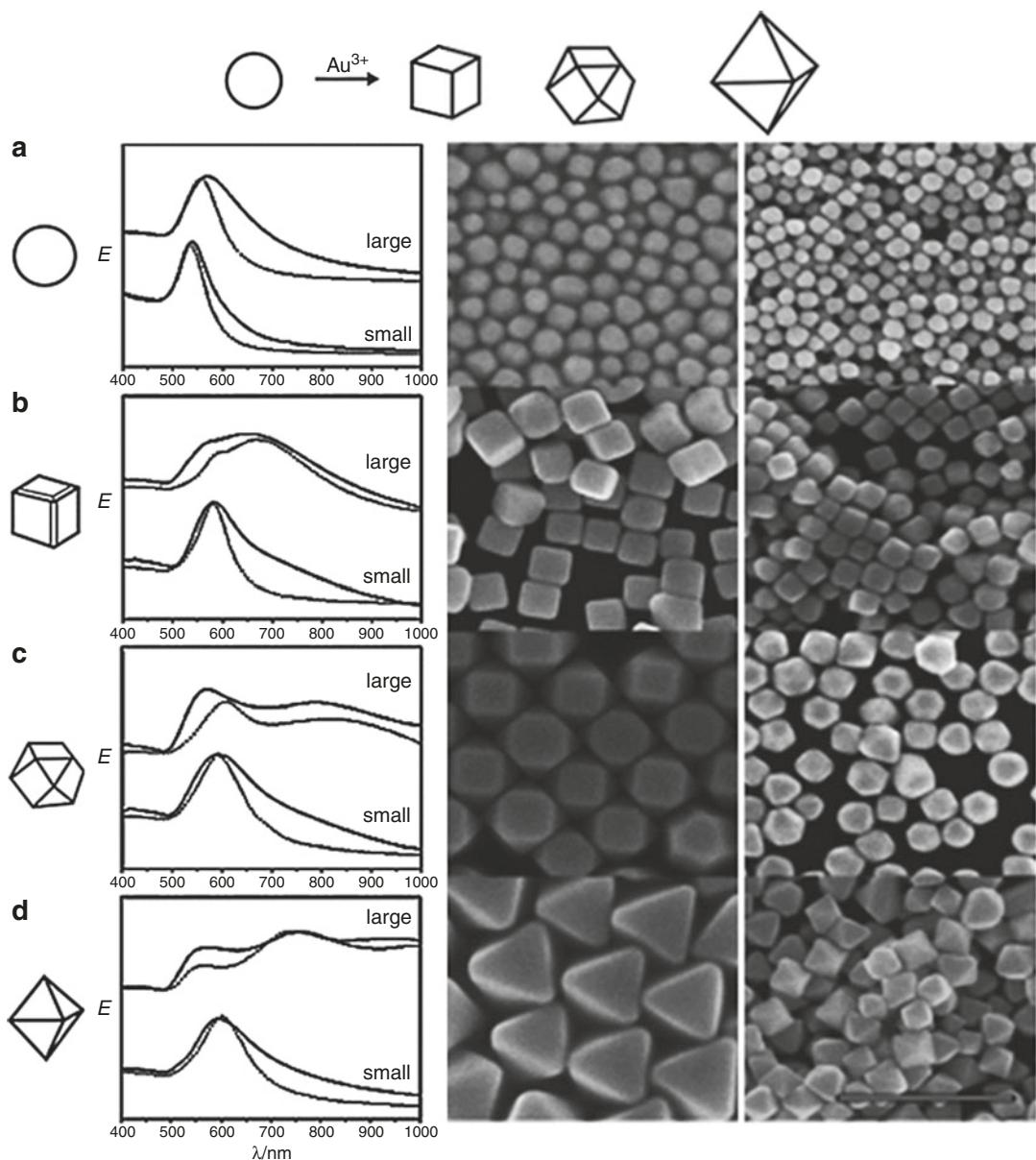
Platonic and quasiplatonic nanocrystals with multiple faces and vertices have been successfully synthesized in relative high yields using both aqueous and nonaqueous approaches [4, 5, 14, 15]. In the method developed by Sau et al. [14] platonic nanoparticles are synthesized in a seeded-growth procedure in aqueous solutions at room temperature. A fine control of nanoparticle shape is achieved by systematic variation of experimental parameters. The experimental procedures involve addition of an appropriate quantity of the presynthesized gold seeds to aqueous growth solutions containing desired quantities of CTAB, HAuCl₄, ascorbic acid, and in some cases a small quantity of silver ions. The morphology and dimension of the gold nanoparticles depend on the concentrations of the seed particles and CTAB, in addition to the reactants (Au³⁺ and ascorbic acid).

The synthesis of platonic gold nanoparticles in nonaqueous media is mainly based on the so-called polyol process. This process involves the thermal reduction of a gold salt in an organic solvent with a relatively high boiling point such as poly(ethylene glycol) or *N,N*-dimethylformamide (DMF) in the presence of a polymeric stabilizer, usually poly(vinyl pyrrolidone). The ratio of gold salt to PVP is an important parameter to control the shape of the polyhedral particles. Tetrahedral particles are formed at higher gold/PVP ratios and icosahedrons at lower ratios. However, the presence of silver ions induces the formation of uniform gold hexahedrons, i.e., nanocubes. Octahedral, truncated octahedral, cubooctahedral, and cubic particles (Fig. 3) were synthesized by Seo et al. by conversion of cubooctahedral particles into different-shaped particles with the aid of silver ions, which seems to be related to selective blocking of the growth along certain crystallographic facets [15]. Also, gold

nanorods, synthesized in aqueous media, can be completely reshaped into perfect, single-crystalline octahedrons in a DMF-PVP solution [4]. Besides reshaping and interconversion methods, faceted seeds can be also employed to grow polyhedral nanoparticles, as in the case of synthesizing very regular and nearly monodisperse decahedrons from preformed pentatwinned gold nanoparticle seeds. The synthesis methods for different platonic nanoparticles along with their chemical and physical properties are summarized in Table 1.

Gold Nanoplates

Other very interesting types of anisotropic nanoparticles are triangular or hexagonal gold nanoplates (Fig. 4), which are highly attractive for a number of applications such as optical biosensing and surface-enhanced Raman spectroscopy (SERS), due to their sharp edges, which lead to high local electric field gradients under illumination [4, 5]. The generation of gold nanoplates is often observed during modifications employed for the synthesis of gold nanorods via the seeded-growth method. Small changes in the synthetic procedure, such as an increase in the pH or surfactant concentration, or the addition of extra halide ions, can lead to the formation of planar nanostructures. To date, all current procedures result in rather low yields (40–65%) compared to the synthesis of gold nanorods ($\sim 99\%$); hence, purification steps are needed in order to eliminate isotropic and nonplanar structures. In the last few years, other techniques rather than seeded-growth or polyol methods have been employed to synthesize nanoplates. Very recently, a more facile synthesis procedure for the fabrication of gold nano- and microplates was reported by Lin et al. [16]. They used a single tree-type multiple-head surfactant, bis (amidoethylcarbamoylethyl) octadecylamine (C₁₈N₃), which functions as both the reducing and capping agent in the reaction system. The triangle and hexagonal plate, polyhedron, and sphere morphology of the gold nanoparticles could be easily controlled simply by changing the molar ratio of C₁₈N₃/HAuCl₄ (Fig. 4).



Synthesis of Gold Nanoparticles, Fig. 3 Size and shape control of Au polyhedral nanocrystals from spherical seeds. UV–Vis extinction spectra (solid lines) and results of DDA calculations (dotted lines) of large (**a**) and small spherical seeds (**a**), large (edge length 116 nm) and small (67 nm) cubes (**b**), large (edge length 122 nm) and small (54 nm) cuboctahedrons (**c**), and large (edge length

236 nm) and small (88 nm) octahedrons (**d**). SEM images (on the right, from top to bottom) of spherical seeds (818 and 495 nm), cubes (11,611 and 676 nm), cuboctahedrons (12,213 and 548 nm), and octahedrons (23,619 and 8,810 nm). The bar represents 500 nm (Reprinted with permission from Ref. [15]. Copyright 2004 Wiley-VCH)

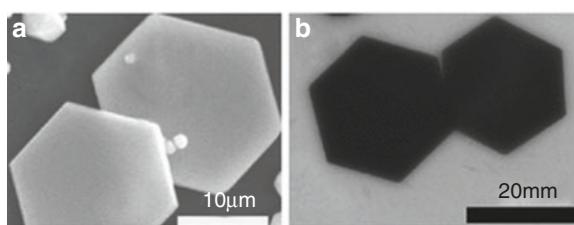
Branched NPs

Coming from symmetric isotropic and anisotropic gold nanoparticles, another important group of gold particles are asymmetric nanoparticles, the

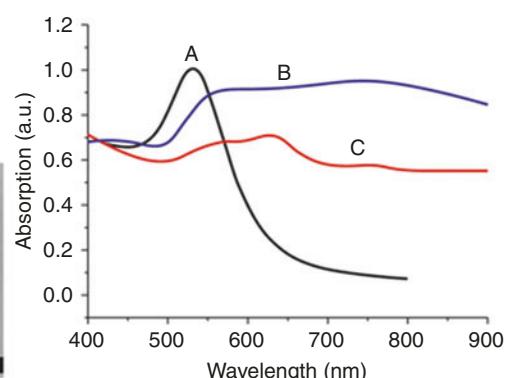
so-called branched nanoparticles or nanostars (Fig. 5) [6]. Although these obviously more complex structures are not yet well understood, they generate interest because of their sharp edges and

Synthesis of Gold Nanoparticles, Table 1 An overview of various synthesis methods of anisotropic gold nanoparticles, with their physical and chemical properties

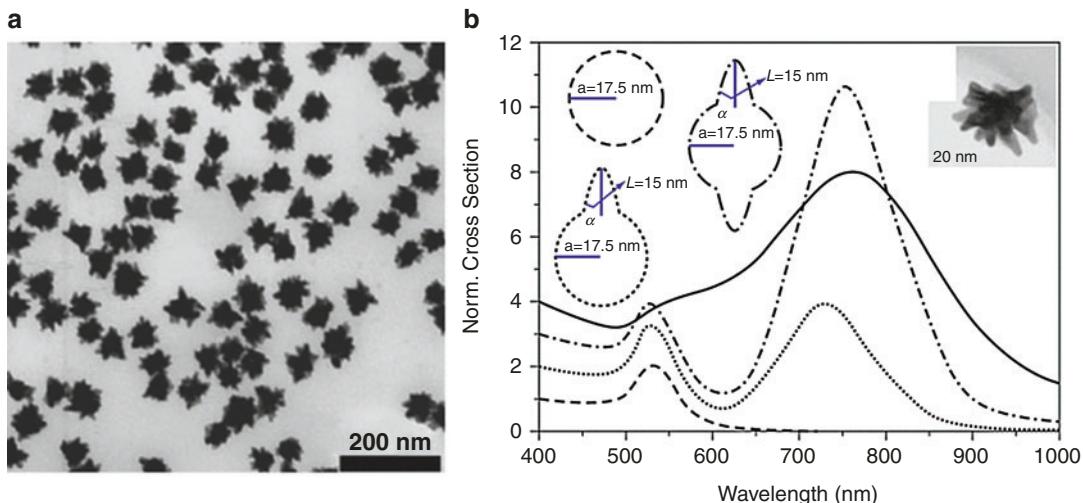
Particle shape	Synthesis method	Max. leng. (nm)	SPR max. (nm)	Synthesis medium	Reduction agent	Capping agent
Rods [18]	Seeded growth	≤ 500 nm	$\lambda_{\text{max tr}}: 520\text{--}530$	H_2O	Ascorbic acid	CTAB
			$\lambda_{\text{max long}}: 750\text{--}800$			
Wires [12]	Seeded growth	$\leq 10,000$	$\lambda_{\text{max tr}}: \sim 540$	H_2O	Ascorbic acid	CTAB
			$\lambda_{\text{max long}}: > 1,200$ nm			
Wires [13]	Seedless growth	$\geq 3,500$	$\lambda_{\text{max tr}}: -$	Oleylamine	Oleylamine	Oleylamine
			$\lambda_{\text{max long}}: > 1,200$ nm			
Tetrahedra [19]	Polyol	~ 210	~ 626 and ~ 950	Ethylene glycol	Ethylene glycol	PVP
Cubes [14]	Seeded growth	~ 90	~ 555	H_2O	Ascorbic acid	CTAB
Cubes [19]	Polyol	~ 150	~ 621	Ethylene glycol	Ethylene glycol	PVP
Octahedra [20]	Polyol	20–400	530–1,100	Ethylene glycol	Ethylene glycol	PDDA
Decahedra [21]	Seeded growth	36–65	600–700	$\text{H}_2\text{O}/\text{DMF}$	PVP	PVP
Decahedra [22]	Polyol	48–88	600–700	Diethylene glycol	Diethylene glycol	PVP
Icosahedra [19]	Polyol	230	~ 613 and ~ 950	Ethylene glycol	Ethylene glycol	PVP
Triangles [23]	Seeded growth	90–210	750–1,300	$\text{H}_2\text{O}/\text{DMF}$	Ascorbic acid	CTAB
Hexagons [16]	Oligoamine surfactant	20–15,000	550–900 broad	H_2O	Oligoamine surfactant	Oligoamine surfactant
Stars [14]	Seeded growth	~ 350	Broad	H_2O	Ascorbic acid	CTAB
Stars [17]	Nonaqueous seeded growth	60–80	Broad	DMF	PVP	PVP



Synthesis of Gold Nanoparticles, Fig. 4 Influence of concentration of C18N3 on the size of gold nanoplates in standard conditions. SEM and TEM images of gold nano- and microplates obtained at different concentrations of C18N3, 2 mM (a) and 0.1 mM (b). On the right are



shown the UV–Vis absorption spectra of gold nanospheres (A), gold decahedrons (B), and gold microplates (C) (Reprinted with permission from Ref. [16]. Copyright 2010 American Chemical Society)



Synthesis of Gold Nanoparticles,

Fig. 5 Representative transmission electron microscopy (TEM) image of Au nanostars (a), experimental (solid line) and calculated (dashed line) absorption spectra for gold

nanostars; the inset shows a TEM image of a single nanostar (b) (Reprinted with permission from Ref. [17]. Copyright IOP Publishing 2008)

the correspondingly high localization of surface plasmon modes. This makes them potential candidates for a number of applications, including SERS-based detection and analysis, as well as the case of nanoplates. Sau et al. proposed the synthesis of branched nanoparticles via the seeded-growth method in the presence of silver ions by varying the seed-to-gold salt ratio and the amount of reducing agent in order to increase the rate of gold ion reduction and thereby induce branching (see Table 1). However, an easier way to synthesize gold nanostars comprises the use of a nonaqueous approach [6, 17]. Branched “nanostars” were synthesized in a concentrated solution of PVP in DMF at room temperature, where PVP acts as a stabilizer and reductant. Nanostars were not only obtained in a seed-mediated growth process but could also be synthesized in a seedless growth process. The fascinating aspect of these particles is that each tip is a single crystal, so the challenge remains to understand the underlying growth mechanism of the selective tip growth on the seed particle’s surface.

In summary, the most relevant synthetic approaches mentioned above for the preparation of different anisotropic nanoparticles are

presented in Table 1. This table offers the opportunity to take a look at the wide range of synthetic possibilities for the production of anisotropic nanoparticles but at the same time gives a good idea of the physical and chemical properties as well as limitations of these synthetic approaches for the production of particular morphologies with specific sizes. That is why further synthetic research efforts are still needed.

Cross-References

- [Cylindrical Gold Nanoparticles](#)
- [Gold Nanorods](#)
- [Hollow Gold Nanoshells](#)
- [Hollow Gold Nanospheres](#)
- [LSPR in Plasmonic Nanostructures: Theoretical Study with Application to Sensor Design](#)
- [Nanobiosensors](#)
- [Nanocluster](#)
- [Nanocrystalline Materials](#)
- [Nanoencapsulation](#)
- [Nanofabrication](#)
- [Nanoparticle](#)
- [Nanoparticles](#)
- [Nanophotonic Structures for Biosensing](#)

- Nanoplasmonics Involved in Localized Photopolymerization
- Nanorods
- Nano-Sized Particle

References

1. Eustis, S.E.-S., El-Sayed, M.A.: Why gold nanoparticles are more precious than pretty gold: noble metal surface plasmon resonance and its enhancement of the radiative and nonradiative properties of nanocrystals of different shapes. *Chem. Soc. Rev.* **35**(3), 209–217 (2006)
2. Astruc, D., Daniel, M.C.: Gold nanoparticles: assembly, supramolecular chemistry, quantum-size-related properties, and applications toward biology, catalysis, and nanotechnology. *Chem. Rev.* **104**(1), 293–346 (2004)
3. Perez-Juste, J., et al.: Gold nanorods: synthesis, characterization and applications. *Coord. Chem. Rev.* **249** (17–18), 1870–1901 (2005)
4. Grzelczak, M., et al.: Shape control in gold nanoparticle synthesis. *Chem. Soc. Rev.* **37**(9), 1783–1791 (2008)
5. Treguer-Delapierre, M., et al.: Synthesis of non-spherical gold nanoparticles. *Gold Bull.* **41**(2), 195–207 (2008)
6. Guerrero-Martínez, A., et al.: Nanostars shine bright for you: colloidal synthesis, properties and applications of branched metallic nanoparticles. *Curr. Opin. Colloid Interface Sci.* **16**(2), 118–127 (2011)
7. Chow, P.E. (ed.): *Gold Nanoparticles: Properties, Characterization and Fabrication Nanotechnology Science and Technology*. Nova, New York (2010)
8. Feldheim, D.L. (ed.): *Metal Nanoparticles: Synthesis Characterization and Applications*. CRC Press, Boca Raton (2001)
9. Jana, N.R., Gearheart, L., Murphy, C.J.: Evidence for seed-mediated nucleation in the chemical reduction of gold salts to gold nanoparticles. *Chem. Mater.* **13**(7), 2313–2322 (2001)
10. Oldenburg, S.J., et al.: Nanoengineering of optical resonances. *Chem. Phys. Lett.* **288**(2–4), 243–247 (1998)
11. Kah, J.C.Y., et al.: Synthesis of gold nanoshells based on the deposition-precipitation process. *Gold Bull.* **41**(1), 23–36 (2008)
12. Kim, F., et al.: Chemical synthesis of gold nanowires in acidic solutions. *J. Am. Chem. Soc.* **130**(44), 14442–14443 (2008)
13. Pazos-Perez, N., et al.: Synthesis of flexible, ultrathin gold nanowires in organic media. *Langmuir* **24**(17), 9855–9860 (2008)
14. Sau, T.K., Murphy, C.J.: Room temperature, high-yield synthesis of multiple shapes of gold nanoparticles in aqueous solution. *J. Am. Chem. Soc.* **126**(28), 8648–8649 (2004)
15. Seo, D., et al.: Directed surface overgrowth and morphology control of polyhedral gold nanocrystals. *Angew. Chem. Int. Ed.* **47**(4), 763–767 (2008)
16. Lin, G.H., et al.: A simple synthesis method for gold nano- and microplate fabrication using a tree-type multiple-amine head surfactant. *Cryst. Growth Des.* **10**(3), 1118–1123 (2010)
17. Kumar, P.S., et al.: High-yield synthesis and optical response of gold nanostars. *Nanotechnology* **19**(1), 015606 (2008)
18. Liu, M.Z., Guyot-Sionnest, P.: Mechanism of silver(I)-assisted growth of gold nanorods and bipyramids. *J. Phys. Chem. B* **109**(47), 22192–22200 (2005)
19. Kim, F., et al.: Platonic gold nanocrystals. *Angew. Chem. Int. Ed.* **43**(28), 3673–3677 (2004)
20. Li, C.C., et al.: A facile polyol route to uniform gold octahedra with tailororable size and their optical properties. *ACS Nano* **2**(9), 1760–1769 (2008)
21. Sanchez-Iglesias, A., et al.: Synthesis and optical properties of gold nanodecahedra with size control. *Adv. Mater.* **18**(19), 2529–2534 (2006)
22. Seo, D., et al.: Shape adjustment between multiply twinned and single-crystalline polyhedral gold nanocrystals: decahedra, icosahedra, and truncated tetrahedra. *J. Phys. Chem. C* **112**(7), 2469–2475 (2008)
23. Millstone, J.E., et al.: Observation of a quadrupole plasmon mode for a colloidal solution of gold nanoprisms. *J. Am. Chem. Soc.* **127**(15), 5312–5313 (2005)
24. Rodriguez-Fernandez, J., et al.: Seeded growth of sub-micron Au colloids with quadrupole plasmon resonance modes. *Langmuir* **22**(16), 7007–7010 (2006)
25. Perez-Juste, J., Correa-Duarte, M.A., Liz-Marzan, L.M.: Silica gels with tailored, gold nanorod-driven optical functionalities. *Appl. Surf. Sci.* **226**(1–3), 137–143 (2004)

S

Synthesis of Graphene

Swastik Kar¹ and Saikat Talapatra²

¹Department of Physics, Northeastern University, Boston, MA, USA

²Department of Physics, Southern Illinois University Carbondale, Carbondale, IL, USA

Synonyms

Fabrication of graphene or graphene fabrication; Graphene synthesis; How to synthesize/make/manufacture/fabricate/produce graphene; Making graphene; Manufacturing graphene or graphene

manufacturing; Production of graphene or graphene production

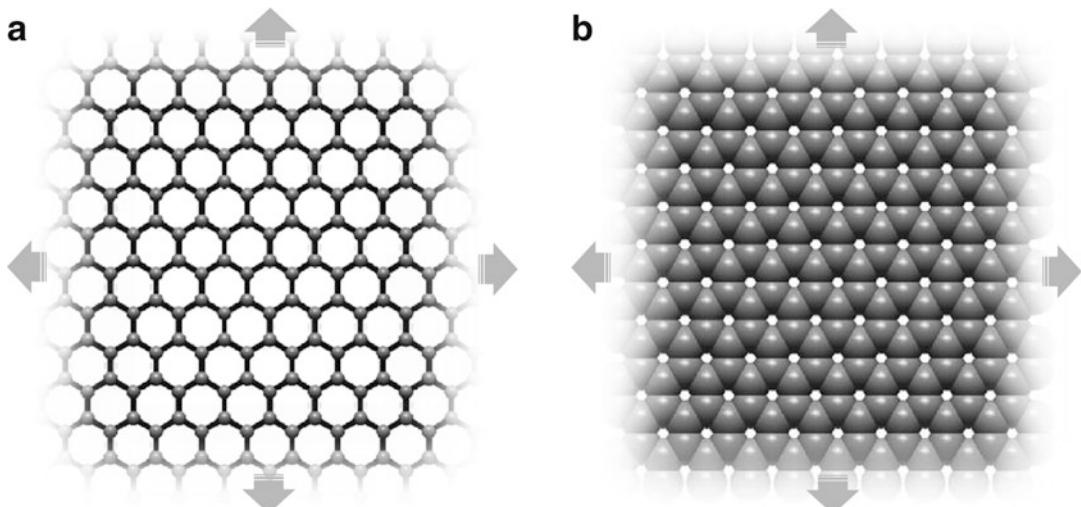
Definition

► **Graphene** (or monolayer graphene or single-layer graphene) is a single-atom-thick, quasi-infinite, sp^2 -hybridized allotrope of carbon in which the atoms are packed in a planar honeycomb crystal lattice (see Fig. 1). It can be visualized as a single sheet of graphite and its lattice structure is related to those of ► **fullerenes for drug delivery** and ► **carbon nanotubes**. Few-layered graphene, multilayered graphene, or multigraphene refers to a few layers of graphene stacked (with weak interlayer attraction) in a manner similar to graphite. Figure 1a,b is a schematic representation of a sheet of single-layer graphene.

Overview

Compiling a review on the diverse methods available for the synthesis of graphene is a daunting task, chiefly due to the explosive intensity at which this field has grown over the past few years. Numerous techniques for graphene

synthesis have been developed, based on both conventional and innovative techniques, and the graphene thus produced can have vastly different properties. Quite often, this diversity stems from a need for the production of graphene that is tailored for a broad variety of experiment-specific and application-specific research. It is quite impossible to enlist all these methods into a few categories, and this entry is not an attempt to individually address all of them. In general, this work attempts to track and identify some of the most popular, novel, and/or promising methods for graphene synthesis. An attempt has been made to categorize these methods into groups. Since the search for new and/or improved methods for graphene synthesis is in its ascent, the material presented here will not reflect published reports beyond May of 2011. In most cases, generic descriptions of the synthesis conditions have been presented. The aim is to provide the reader with a starting point rather than a complete recipe in each case. It should be noted that experimental conditions and parameters presented here represent typical synthesis conditions as reported in the literature, since what constitutes the “best practice” is both subjective and continuously evolving. It should also be noted that this entry does not deal with the many electronic, optical, thermal, mechanical, or



Synthesis of Graphene, Fig. 1 Schematic representation of carbon atoms in a graphene lattice: (a) a ball-and-stick model and (b) a space-filling model. The arrows

indicate that the sheets are quasi-infinitely extended in all directions. The edge carbon atoms are usually hydrogen terminated

other fascinating properties of graphene, for which excellent review articles already exist in the literature.

Historical Perspective

The concept of graphene has been around for quite a long time, and the descriptive presence of this material appears in a variety of reports in a number of languages. For example, “very thin layers of graphite” including “single foils of carbon” were reported as early as 1962, prepared through a partial reduction of graphite oxide [1]. Monolayer graphite/carbon was identified to form at elevated temperatures on surfaces of Pt in 1969 [2], C-doped Ni (111) crystals in 1979 [3], and other transition metal surfaces [4]. The term “graphene” was already in use in the early 1990s [5], to describe not only monolayer graphite but also the structure of carbon fibers and fullerenes. An earnest endeavor to obtain isolated high-quality graphene from graphite started in 1999, using mechanical exfoliation of patterned graphite pillars [6, 7]. This approach matured in 2004 [8], the same year that synthesis of graphene was also reported by a high-temperature thermal annealing of SiC [9]. These two methods opened up the floodgate of graphene research and catapulted it into becoming the millennium material. In the following paragraphs, a flavor of these and some of the other popular graphene synthesis methods that have been developed since then has been captured.

Recent Advances in Graphene Synthesis

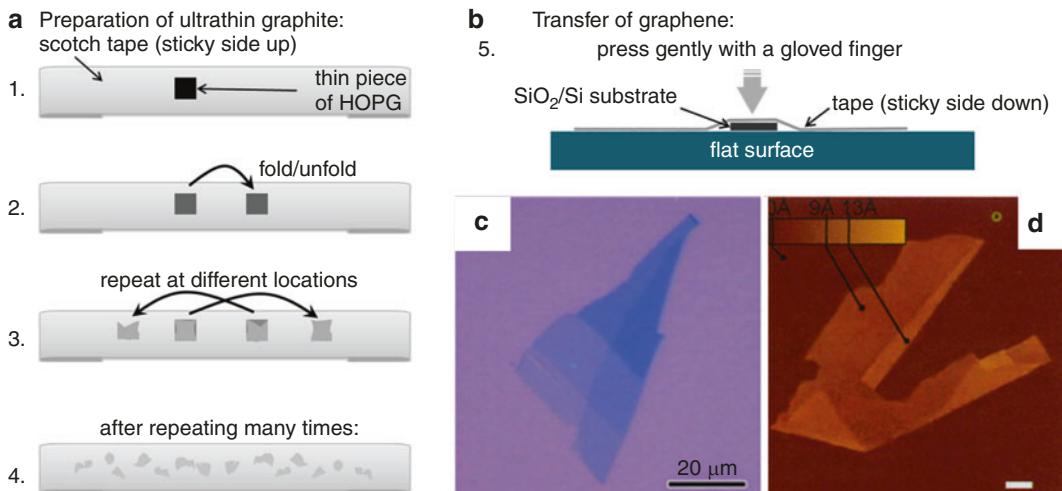
Micromechanical Exfoliation of Graphene from Graphite

Strictly not a synthesis process, the micromechanical exfoliation of graphene (both monolayer and few-layer graphene) from highly oriented pyrolytic graphite (HOPG), as reported by Novoselov et al., enabled, for the first time, a direct measurement of its unique 2D electronic properties. This ended the decades-long debate whether pure monatomic two-dimensional crystals

can indeed be stable in an isolated state [10, 11]. The few tens-of-microns-sized graphene and few-layered graphene flakes could be easily transferred onto SiO₂/Si and other substrates and enabled researchers to perform a wide variety of electronic, optical, thermal, and mechanical property measurements. In their first report [8], Novoselov et al. described an elaborate process that started from millimeter-thick platelets of HOPG and involved patterning, etching, photore sist, an adhesive-tape peel-off, and a subsequent ultrasonic rinse. A simplified prescription for obtaining graphite was reported by the same group a year later, whereby a freshly cleaved HOPG surface was mechanically rubbed against any solid surface. This invariably left behind debris of crystallites on the surface, and among them, one can usually find single- and few-layered graphene [12]. Owing to the simplicity of this method, a variety of alternatives have become popular, and Fig. 2a, b describes a protocol that simply uses an adhesive tape, which is favored by many. This method has become known as the scotch tape method. As a receiving substrate, it is helpful to use a Si wafer that has ~300-nm-thick oxide layer: graphene itself is nearly transparent, but using this oxide thickness enables one to “see” graphene even using an optical microscope due to the suitable amount of interferometric phase contrasting, as seen in Fig. 2c [13]. Figure 2d is an ► atomic force microscopy (AFM) image of a typical flake of graphene, showing the graphene layer thicknesses in flat and folded regions [12].

Epitaxial Growth of Graphene on Silicon Carbide

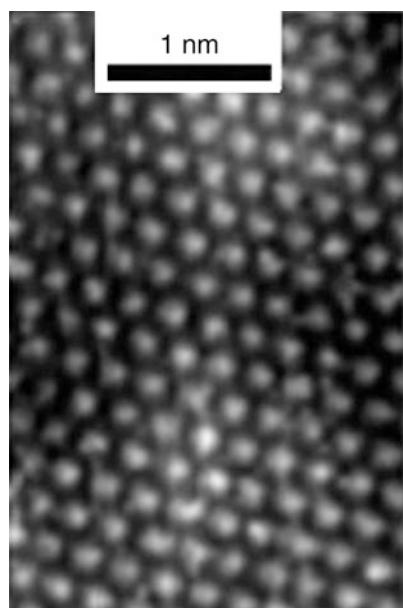
In 2004, another technique for graphene synthesis was reported by Berger et al. [9]. In this method, graphene crystals were produced on the Si-terminated (0001) face of single-crystal 6H-SiC by thermal desorption of Si. The SiC surfaces were initially prepared by oxidation or H₂ etching. The oxide was then removed by electron bombardment-assisted heating in ultrahigh vacuum ($\sim 10^{-10}$ Torr) to 1,000 °C (repeating the oxidation/deoxidation cycle improves the surface quality, and the best initial surface quality was obtained with H₂ etching). The deoxidized



Synthesis of Graphene, Fig. 2 A simple recipe for obtaining single- and few-layered graphene crystals using the micromechanical cleavage of graphite. (a) 1 Attach a very thin piece of HOPG onto a scotch tape. 2–4 Fold and unfold as shown to obtain ultrathin (almost invisible) flakes of graphite on the tape. (b) Select a region with very thin flakes and transfer on a suitable substrate (e.g., SiO₂/Si) as shown. Carefully remove the tape from the substrate. (c)

Typical image of a few-layered flake with monolayer extensions as seen under an optical microscope (Reprinted with permission from Ref. [13]; copyright © 2010 Elsevier). (d) A typical AFM image of graphene flakes (scale bar = 1 μm) (Reprinted with permission from Ref. [12]; copyright © 2005 National Academy of Sciences, U.S.A.)

samples were heated to temperatures ranging from 1,250 °C to 1,450 °C for 1–20 min, during which time, thin graphite layers were formed. Figure 3 shows a scanning tunneling microscope (STM) image of graphene produced this way by heating SiC for 8 min at 1,400 °C. Over the years, this process has been fine-tuned so as to enable one to controllably produce monolayer graphene on SiC substrates [14]. It turns out that graphenes grown from the SiC(000\$ \$ \bar{1} \$ \$) (C-terminated) surface are of higher quality than those grown on SiC(0001) [15]. Growth on the C face renders graphene with domain sizes more than three times larger than those grown on the Si face and with significantly reduced disorder. In 2009, it was reported that an ex situ graphitization of Si-terminated SiC (0001) in an argon atmosphere of about 1 bar produces monolayer graphene films with much larger domain sizes than those previously attained. This method for the direct synthesis of large-domain graphene (~10 s of microns in length, with ~micron-sized widths that were limited by the step sizes of the underlying substrate)



Synthesis of Graphene, Fig. 3 Atomically resolved STM image of a graphene sample grown on SiC(0001) at 1,400 °C for 8 min (Reprinted with permission from Ref. [9]; copyright © 2004 American Chemical Society)

on an insulating underlayer formed an important step toward realizing graphene-based nanoelectronics on a semiconductor-processable substrate.

In the initial years, most samples were produced using one of the two abovementioned methods, which produced high-quality samples of graphene appropriate for fundamental studies such as electric- and magnetic-field-modulated electronic transport, ARPES studies, Raman studies, and a variety of related experiments. At the same time, as it became evident that there are several other important properties of graphene including their gigantic specific surface area, extreme mechanical strength values, and other thermal and optical properties, several other methods have been developed for large-scale production of graphene. The large-scale production can be categorized into two directions. One route is to develop methods for mass production, often using liquid-phase exfoliation of graphene from graphite, while the other is headed toward wafer-scale production of single- or few-layered graphene, targeted toward electrical, electromechanical, and optoelectronic applications.

Liquid-Phase Exfoliation

Reduced Graphene Oxide and Chemically Modified Graphene

A significant amount of effort has gone into obtaining graphene from the exfoliation of graphite oxide into graphene oxide (GO), followed by a reduction step [16]. The resultant material is more appropriately termed reduced graphene oxide (rGO), and it belongs to the broader category of chemically modified graphenes. Typically, graphite is oxidized using techniques based on Hummer's method [17]. GO, as produced by Hummers' method, is composed of functionalized graphene sheets decorated by strongly bound oxidative debris, which acts as a surfactant to stabilize aqueous GO suspensions [18]. The oxide can hence be dispersed in water using an ultrasonic bath till it becomes a clear-colored liquid with no visible particulates. The parent graphite oxide and the dispersed graphene oxide are both insulators and hence are unsuitable for any applications

where electrical conductivity is important. One approach to overcome this problem is a liquid- or gas-phase reduction of the oxidized carbon. Among reducing agents, the use of hydrazine hydrate has become quite popular, and this can be added to the aqueous GO dispersion and heated in an oil bath at 100 °C for a day, during which time, the reduced graphene oxide gradually precipitates out as a black solid. This precipitate is a high-surface-area material which can be separated out using a simple vacuum filtration technique that leaves behind a thick film of rGO. Such films can be transferred onto other substrates in the form of large-area ultrathin films of reduced graphene oxide, which are useful for the development of transparent and flexible electronic materials [19] (see also ► **Flexible Electronics**) and other applications [20, 21].

Although the reduction process restores its conductivity by several orders of magnitude compared to that of GO [22], the oxidation of graphene seems to be quite robust and can only be partially removed. Solid-state ^{13}C NMR spectroscopy of these materials suggests that the sp^2 -bonded carbon network of graphite is strongly disrupted, and a significant fraction of this carbon network is bonded to hydroxyl groups or participates in epoxide groups [23], with some carboxylic or carbonyl groups present at the edges. Further works have shown the presence of five- and six-membered ring lactols [24]. As a result, this material has a poor electrical conductivity compared to that of pristine graphene and does not display most of the exciting quantum properties of pure graphene. These groups can be significantly removed by using a sodium borohydride and sulfuric acid treatment, followed by thermal annealing [24], which is effective in restoring the π -conjugated structure and to a significant extent the conductive nature of the graphene materials. While the effort to restore the electronic properties of rGO to a level similar to that of pristine graphene has yet to be reported, rGO-related materials can have several important properties of their own, including large specific surface area and other mechanical properties that enable them to be used as various kinds of energy storage devices (► **Nanomaterials for Electrical Energy**

Storage Devices), sensors, and composite structures.

Liquid-Phase Exfoliation of Graphene Directly from Graphite

In some applications, where it is important to have large volumes of *chemically unmodified* graphene, a different approach can be adopted, involving liquid-phase exfoliation of graphene directly from graphite. The idea of liquid-phase exfoliation is essentially a combination of the two methods – mechanical exfoliation from graphite and the liquid-phase environment that is usually applied to GO. In case of GO, the process is pretty straightforward, owing to the weaker interplanar interactions and the strong hydrophilic nature of the individual graphene sheets, and simple ultrasonication of GO causes it to easily disperse stably in water. In contrast, graphene sheets in graphite interact more strongly and, being hydrophobic, do not disperse readily into aqueous media.

Early efforts to exfoliate graphite were mostly through the intercalation of graphite using a range of compounds [25]. With the resurging interest in graphene, this idea has been taken forward to produce large-scale exfoliation of graphene from commercially available expandable graphite [26]. Expandable graphite can be prepared by chemical intercalation of sulfuric acid and nitric acid. Upon heating, the volatile gaseous species released from the intercalant help exfoliate graphite very rapidly into multilayer graphene. This product is then reintercalated with oleum (fuming sulfuric acid with 20 % free SO₃), and tetrabutylammonium hydroxide (TBA, 40 % solution in water) is inserted into the oleum-intercalated graphite. The TBA-inserted oleum-intercalated graphite is sonicated in a *N,N*-dimethylformamide (DMF) solution of 1,2-distearoyl-sn-glycero-3-phosphoethanolamine-*N*-[methoxy(polyethyleneglycol)-5000](DSPE-mPEG) for 60 min to form a homogeneous suspension. A process of centrifugation removes large pieces of material from the supernatant and results in large amounts of graphene sheets suspended in DMF. The exfoliation-reintercalation-expansion of graphite can produce

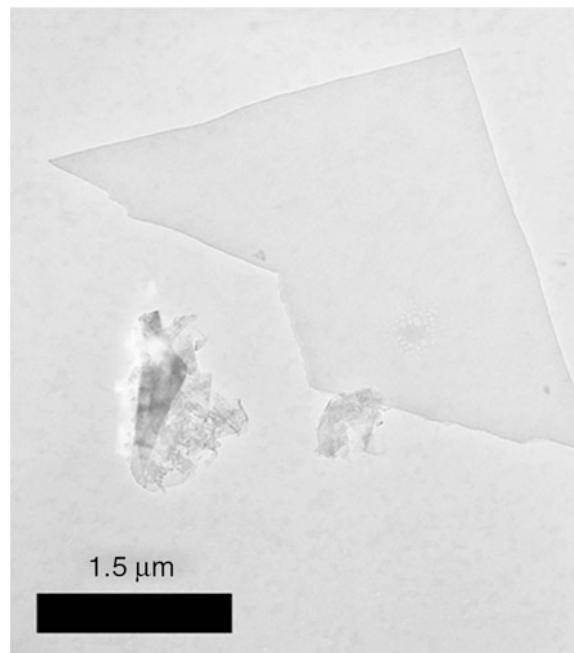
high-quality single-layer graphene sheets stably suspended in organic solvents and possessing high electrical conductance and can be collected as large transparent conducting films by Langmuir-Blodgett assembly in a layer-by-layer manner.

In a more simple approach, stable dispersions up to ~0.01 mg/ml can be directly exfoliated from graphite by ultrasonication in organic solvents such as *N*-methyl pyrrolidone [27]. This method is physicochemically possible because the energy required to exfoliate graphene is balanced by the solvent-graphene interaction for solvents whose surface energies match that of graphene. Typically, graphite is dispersed in a relevant solvent by sonicating in a low-power ultrasonic bath for 30 min. The resultant dispersion is then centrifuged for 90 min at 500 rpm and decanted by pipetting off the top half of the dispersion. The decanted product usually contains a mixture of single- and few-layer graphene flakes that can be made into thin films by passing them through nanoporous membranes and drying them off.

A different method of obtaining stable dispersions in a liquid medium was reported by An et al. [28]. In this method, 1-pyrenecarboxylic acid (PCA) was used as a method to “wedge” out graphene sheets from graphite in suitably polarity-controlled combination of media to give rise to noncovalently functionalized graphene sheets that were stably dispersed in water. In a controlled medium, initially, PCA serves as a “molecular wedge” that cleaves the individual graphene flakes from the parent graphite pieces and then forms stable polar functional groups on the graphene surface via a noncovalent π-π stacking mechanism that does not disrupt its sp² hybridization. Figure 4 shows a typical TEM (► Transmission Electron Microscopy) image of graphene obtained this way. The hydrophilic -COOH group of PCA facilitates the formation of stable aqueous dispersions of graphene, in a manner similar to that of graphene oxide, but without degrading the sp² structure. The advantage of this method lies in the fact that the final product is stable in water – opening up the possibilities of experiments where an aqueous medium could be important, such as interactions with

Synthesis of Graphene,

Fig. 4 (a) A TEM image of a typical graphene flake exfoliated from graphite using a molecular wedging technique [28]



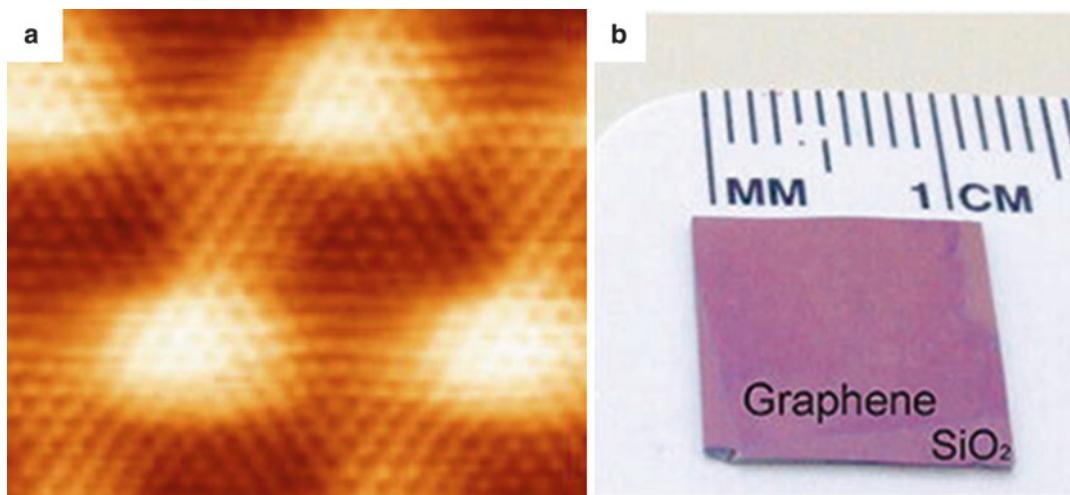
biological samples. At the same time, a simple methanol wash will remove the excess PCA from the surface and the pure graphene sheet precipitates out with time – to be utilized for experiments where pure, unfunctionalized graphene sheets are required. Similar methods have been demonstrated by other groups to obtain dispersions of graphene from graphite in surfactant-water solutions [29], through which significant levels of exfoliation and monolayer formation can be achieved.

Epitaxial and Large-Area Graphene on Metal Surfaces

High-temperature annealing of Pt single crystals produces an atomically thin entity on their surfaces, and this layer has been identified as monolayer graphite back in 1969 through low-energy electron diffraction (LEED) experiments [2]. Since then, it became clear that graphene in this morphology appears either due to the surface desegregation of carbon or through decomposition of hydrocarbons. Graphene has been reported to grow on a variety of metals such as Co, Ni, Ru, Rh, Pd, Ir, and Pt, with varying degrees of grain size, epitaxial alignment with the underlying

metal, defect densities, and surface morphologies [30]. For example, it was shown in 2006 that ethylene, absorbed at room temperature on Ir (111) planes, could be desegregated at 1,450 K to produce graphene flakes of ~100 nm size that covers about 30 % of the surface [31]. More recently, the desegregation method was successfully utilized to graphene on Ru (0001) surfaces, where macroscopic single-crystalline domains with linear dimensions exceeding 200 μm could be achieved [32, 33]. Figure 5a shows a high-resolution STM (► Scanning Tunneling Microscopy) image of graphene grown on Ru surface. By controlling the chamber pressure, growth time, and temperature, millimeter-scale graphene was reported on Ru surfaces using a controlled desegregation process in 2009 [34].

In 2008, the desegregation method received a boost when a low-pressure ► chemical vapor deposition (CVD) technique was utilized to synthesize graphene on an Ir (111) surface in a controlled manner over a range of temperatures from 1,120 to 1,320 K [35]. Scanning tunneling microscopy of the grown structures reveals that graphene prepared using this technique exhibits a continuity of its carbon rows over terraces and



Synthesis of Graphene, Fig. 5 (a) Atomically resolved STM image of the graphene layer grown on Ru (0001) (Reprinted with permission from Ref. [32]; copyright © (2007) by the American Physical Society). (b) A digital photograph of a graphene film grown on a Cu

foil using a CVD method and transferred onto a SiO₂/Si substrate (Reprinted with permission from Ref. [38]; copyright © (2009) by the American Association for the Advancement of Science)

step edges, enabling the possibility of large-scale growth that is not limited by surface features of metals. The thermodynamics of carbon segregation and graphene formation involves a complex interplay of carbon diffusion from the bulk to the surface of the host metal, followed by surface diffusion of C atoms that lead to nucleation and growth. In a typical CVD process, a metal substrate is placed in a flow-type furnace which is heated to a temperature between 750 °C and 1,000 °C in the presence of a mixture of Ar and H₂ at a fixed chamber pressure P. Once the target temperature stabilizes, the flow is maintained for some time (30–60 min) to clean and prepare the surface of the metal. At the end of this preparation time, the carbon source, typically ethylene or methane, is introduced for a fixed time, at the end of which the furnace is allowed to cool down. Once cooled, the surface is found to be covered with single- and multilayered graphene. Factors that appear to control layer thickness and domain sizes of graphene obtained such as temperature, feed-gas concentration, chamber pressure, and time appear to be vastly different for different metals. For example, Yu et al. used CH₄:H₂:Ar = 0.15:1:2 with a total gas flow rate

of 315 sccm and found that the type of graphene obtained on Ni foils depended strongly on the cooling rate, with multilayer graphene obtained only when the cooling rate was controlled to 10 °C/s [36]. A great benefit of this technique is that thin layers of pre-patterned Ni could be utilized to fabricated large-scale patterned growth of graphene films that could be transferred to arbitrary substrates [37].

At elevated temperatures, the solubility of C atoms in Ni is much higher compared to that in Cu, and the resulting graphene formation can have different morphologies for these two metals. Centimeter-scale graphene films were shown to grow rather easily on Cu substrates in 2009 through a similar CVD process [38], with more than 95 % of the surface found to be covered with single-layer graphene that are continuous across copper surface steps and grain boundaries. Figure 5b shows an example of a macroscopic sheet of graphene grown on a Cu foil and transferred onto a SiO₂/Si substrate. In case of Cu, it seems that the cooling rate is unimportant, most probably due to the low solubility of carbon in copper which helps to make the growth process self-limiting. There were no discernible

differences in the graphene morphology when the cooling rate was varied from $>300\text{ }^{\circ}\text{C}/\text{min}$ to about $40\text{ }^{\circ}\text{C}/\text{min}$. The ability to grow high-quality large-scale graphene on a low-cost material (Cu) at such ease has been rapidly taken advantage of to produce graphene at a roll-to-roll capability [39]. This, along with a wet chemical doping and layer-by-layer stacking process, can be used to devise many-layer macroscopic graphene films with sheet resistance at values as low as $\sim 30\text{ }\Omega\text{ }^{-1}$ at $\sim 90\text{ }\%$ transparency, which is superior to commercial transparent electrodes such as indium tin oxides and opens up several new possibilities in the fields of optics, optoelectronics, and photovoltaics.

Other Novel Advances in Graphene Synthesis

In addition to the categories described above, there are numerous innovative methods being reported on a regular basis that are either extensions of the abovementioned categories or are completely new methods that can become significantly useful and/or popular in future. In addition to graphene, there is a whole array of related materials, including doped and functionalized graphene and graphene nanoribbons, each of which is unique and important by its own rights. It is impossible to either enlist all or describe these methods individually within the scope of this entry. Some of the more popular and/or innovative ones are enlisted in Table 1 in order to provide the reader with a starting point. It should be borne in mind that a keyword search of existing databases through available searching tools was employed to obtain these references, and omission of any significant report(s) is purely accidental and inadvertent.

Conclusion

It is evident that within a very short time frame of a few years, the scientific community has witnessed a tremendous advancement in terms of various techniques of graphene synthesis. This advancement has progressed hand-in-hand with a number of extremely important fundamental discoveries related to this novel 2D crystalline

Synthesis of Graphene, Table 1 Some recent, novel, and uncategorized graphene synthesis techniques

Method	Publication details
Solvothermal synthesis	Choucair et al. Nat. Nanotechnol. 4 , 30–33 (2009)
Substrate-free synthesis	Dato et al. Nano Lett. 8 , 2012–2016 (2008)
Electrochemical synthesis	1. Liu et al. Adv Func. Mat. 18 , 1518–1525 (2008) 2. Guo et al. ACS Nano 3 , 2653–2659 (2009)
Microwave plasma enhanced CVD	Malesevic et al. Nanotech. 19 , 305604.1–305604.6 (2008)
Arc discharge exfoliation	Wu et al. ACS Nano 3 , 411–417 (2009)
Microwave-solvothermal synthesis	Murugan et al. Chem. Mat. 21 , 5004–5006 (2009)
Doped graphene	1. Panchokarla et al. Adv Mat 21 , 4726–4730 (2009) 2. Ci et al. Nat. Mater. 9 , 430–435 (2010) 3. Li et al. Carbon 48 , 255–259 (2010)
Graphite exfoliation in ionic liquids	Lu et al. ACS Nano 3 , 2367–2375 (2009)
Hydrophilic and organophilic graphene	Wang et al. Carbon 47 , 1359–1364 (2009)
Graphene nanoribbons	1. Han et al. Phys. Rev. Lett. 98 , 206805.1–206805.4 (2007) 2. Li et al. Science 319 , 1229–1232 (2008) 3. Jiao et al. Nat. Nanotech. 5 , 321–325 (2010)
Reducing sugar	Zhu et al. ACS Nano 4 , 2429–2437 (2010)

material and has also opened up several avenues for groundbreaking technological directions. Despite the rapid success that this field has experienced, the interest in developing better and more cost-effective synthesis methods keeps growing. In the near term, production of high-quality graphene with reproducible electrical as well as mechanical properties will be a challenge. Similarly, techniques and parameters for size-specific, controlled synthesis of graphene such as graphene nanoribbons, graphene quantum dots, etc., need to be well understood. Processes for large-scale synthesis of monolayer graphene (e.g., through chemical exfoliation) need to be perfected in order to

utilize these materials in a variety of bulk applications such as electrodes in electrical energy storage systems, fillers in composite, etc. Overall, the current level of progress as well as the ongoing research efforts in graphene synthesis has been quite promising for certain niche area applications, and the scientific community is quite optimistic that in the future, graphene will eventually emerge as the next ultimate engineering material.

Cross-References

- [Atomic Force Microscopy](#)
- [Carbon Nanotubes](#)
- [Chemical Vapor Deposition \(CVD\)](#)
- [Flexible Electronics](#)
- [Fullerenes for Drug Delivery](#)
- [Graphene](#)
- [Nanomaterials for Electrical Energy Storage Devices](#)
- [Scanning Tunneling Microscopy](#)
- [Transmission Electron Microscopy](#)

References

1. Boehm, H.P., Clauss, A., Fischer, G.O., Hofmann, U.: Das Adsorptionsverhalten sehr dünner Kohlenstofffolien. *Z. Anorg. Allg. Chem.* **316**, 119–127 (1962)
2. May, J.W.: Platinum surface LEED rings. *Surf. Sci.* **17**, 267–270 (1969)
3. Eizenberg, M., Blakely, J.M.: Carbon monolayer phase condensation on Ni(111). *Surf. Sci.* **82**, 228–236 (1979)
4. Aizawa, T., Souda, R., Otani, S., Ishizawa, Y., Oshima, C.: Anomalous bond of monolayer graphite on transition-metal carbide surfaces. *Phys. Rev. Lett.* **64**, 768–771 (1990)
5. Dresselhaus, M.S., Dresselhaus, G., Saito, R.: Carbon fibers based on C60 and their symmetry. *Phys. Rev. B* **45**, 6234–6242 (1992). Aizawa, T., Hwang, Y., Hayami, W., Souda, R., Otani, S., Ishizawa, Y.: Phonon dispersion of monolayer graphite on Pt(111) and NbC surfaces: bond softening and interface structures. *Surf. Sci.* **260**, 311–318 (1992)
6. Lu, X., Yu, M., Huang, H., Ruoff, R.S.: Tailoring graphite with the goal of achieving single sheets. *Nanotechnology* **10**, 269–272 (1999)
7. Lu, X., Huang, H., Nemchuk, N., Ruoff, R.S.: Patterning of highly oriented pyrolytic graphite by oxygen plasma etching. *Appl. Phys. Lett.* **75**, 193–195 (1999)
8. Novoselov, K.S., Geim, A.K., Morozov, S.V., Jiang, D., Zhang, Y., Dubonos, S.V., Grigorieva, I.V., Firsov, A.A.: Electric field effect in atomically thin carbon films. *Science* **306**, 666–669 (2004)
9. Berger, C., Song, Z., Li, T., Li, X., Ogbazghi, A.Y., Feng, R., Dai, Z., Marchenkov, A.N., Conrad, E.H., First, P.N., de Heer, W.A.: Ultrathin epitaxial graphite: 2D electron gas properties and a route toward graphene-based nanoelectronics. *J. Phys. Chem. B* **108**, 19912–19916 (2004)
10. Peierls, R.E.: Quelques propriétés typiques des corps solides. *Ann. Inst. Henri. Poincaré* **5**, 177–222 (1935)
11. Landau, L.D.: Zur Theorie der phasenumwandlungen II. *Phys. Z. Sowjetunion* **11**, 26–35 (1937)
12. Novoselov, K.S., Jiang, D., Schedin, F., Booth, T.J., Khotkevich, V.V., Morozov, S.V., Geim, A.K., Rice, T. M.: Two-dimensional atomic crystals. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 10451–10453 (2005)
13. Soldano, C., Mahmood, A., Dujardin, E.: Production, properties and potential of graphene. *Carbon* **48**, 2127–2150 (2010)
14. Rollings, E., Gweon, G.-H., Zhou, S.Y., Mun, B.S., McChesney, J.L., Hussain, B.S., Fedorov, A.N., First, P.N., de Heer, W.A., Lanzara, A.: Synthesis and characterization of atomically thin graphite films on a silicon carbide substrate. *J. Phys. Chem. Solids* **67**, 2172–2177 (2006)
15. Hass, J., Feng, R., Li, T., Li, X., Zong, Z., de Heer, W. A., First, P.N., Conrad, E.H., Jeffrey, C.A., Berger, C.: Highly ordered graphene for two dimensional electronics. *Appl. Phys. Lett.* **89**, 143106–143108 (2006)
16. Stankovich, S., Dikin, D.A., Piner, R.D., Kohlhaas, K. A., Kleinhammes, A., Jia, Y., Wu, Y., Nguyen, S.T., Ruoff, R.S.: Synthesis of graphene-based nanosheets via chemical reduction of exfoliated graphite oxide. *Carbon* **45**, 1558–1565 (2007)
17. Hummers Jr., W.S., Offeman, R.E.: Preparation of graphitic oxide. *J. Am. Chem. Soc.* **80**, 1339–1339 (1958)
18. Rourke, J.P., Pandey, P.A., Moore, J.J., Bates, M., Kinloch, I.A., Young, R.J., Wilson, N.R.: The real graphene oxide revealed: stripping the oxidative debris from the graphene-like sheets. *Angew. Chem. Int. Ed.* **50**, 3173–3177 (2011)
19. Eda, G., Fanchini, G., Manish Chhowalla, M.: Large-area ultrathin films of reduced graphene oxide as a transparent and flexible electronic material. *Nat. Nanotechnol.* **3**, 270–274 (2008)
20. Gilje, S., Han, S., Wang, M., Wang, K.L., Kaner, R.B.: A chemical route to graphene for device applications. *Nano Lett.* **7**, 3394–3398 (2007)
21. Tung, V.C., Allen, M.J., Yang, Y., Kaner, R.B.: High-throughput solution processing of large-scale graphene. *Nat. Nanotechnol.* **4**, 25–29 (2009)
22. Si, Y., Samulski, E.T.: Synthesis of water soluble graphene. *Nano Lett.* **8**, 1679–1682 (2008)

23. Park, S., Ruoff, R.S.: Chemical methods for the production of graphenes. *Nat. Nanotechnol.* **4**, 217–224 (2009)
24. Gao, W., Alemany, L.B., Ci, L., Ajayan, P.M.: New insights into the structure and reduction of graphite oxide. *Nat. Chem.* **1**, 403–408 (2009)
25. Dresselhaus, M.S., Dresselhaus, G.: Intercalation compounds of graphite. *Adv. Phys.* **30**, 139–326 (1981)
26. Li, X., Zhang, G., Bai, X., Sun, X., Wang, X., Wang, E., Dai, H.: Highly conducting graphene sheets and Langmuir–Blodgett films. *Nat. Nanotechnol.* **3**, 538–542 (2008)
27. Hernandez, Y., Nicolosi, V., Lotya, M., Blighe, F.M., Sun, Z., De, S., McGovern, I.T., Holland, B., Byrne, M., Gun’Ko, Y.K., Boland, J.J., Niraj, P., Duesberg, G., Krishnamurthy, S., Goodhue, R., Hutchison, J., Scardaci, V., Ferrari, A.C., Coleman, J.N.: High-yield production of graphene by liquid-phase exfoliation of graphite. *Nat. Nanotechnol.* **3**, 563–568 (2008)
28. An, X., Simmons, T., Shah, R., Wolfe, C., Lewis, K. M., Washington, M., Nayak, S.K., Talapatra, S., Kar, S.: Stable aqueous dispersions of noncovalently functionalized graphene from graphite and their multifunctional high-performance applications. *Nano Lett.* **10**, 4295–4301 (2010)
29. Lotya, M., Hernandez, Y., King, P.J., Smith, R.J., Nicolosi, V., Karlsson, L.S., Blighe, F.M., De, S., Wang, Z., McGovern, I.T., Duesberg, G.S., Coleman, J.N.: Liquid phase production of graphene by exfoliation of graphite in surfactant/water solutions. *J. Am. Chem. Soc.* **131**, 3611–3620 (2009)
30. Wintterlin, J., Bocquet, M.-L.: Graphene on metal surfaces. *Surf. Sci.* **603**, 1841–1852 (2009)
31. N’Diaye, A.T., Bleikamp, S., Feibelman, P.J., Michely, T.: Two-dimensional Ir cluster lattice on a graphene moire on Ir(111). *Phys. Rev. Lett.* **97**, 215501.1–215501.4 (2006)
32. Marchini, S., Gunther, S., Wintterlin, J.: Scanning tunneling microscopy of graphene on Ru(0001). *Phys. Rev. B* **76**, 075429.1–075429.9 (2007)
33. Sutter, P.W., Flege, J.-I., Sutter, E.A.: Epitaxial graphene on ruthenium. *Nat. Mater.* **7**, 406–411 (2008)
34. Pan, Y., Zhang, H.G., Shi, D.X., Sun, J., Du, S., Liu, F., Gao, H.-J.: Highly ordered, millimeter-scale, continuous, single-crystalline graphene monolayer formed on Ru (0001). *Adv. Mater.* **21**, 2777–2780 (2009)
35. Coraux, J., N’Diaye, A.T., Busse, C., Michely, T.: Structural coherency of graphene on Ir(111). *Nano Lett.* **8**, 565–570 (2008)
36. Yu, Q., Lian, J., Siriponglert, S., Li, H., Chen, Y.P., Pei, S.-S.: Graphene segregated on Ni surfaces and transferred to insulators. *Appl. Phys. Lett.* **93**, 113103–113105 (2008)
37. Kim, K.S., Zhao, Y., Jang, H., Lee, S.Y., Kim, J.M., Kim, K.S., Ahn, J.-H., Kim, P., Choi, J.-Y., Hong, B. H.: Large-scale pattern growth of graphene films for stretchable transparent electrodes. *Nature* **457**, 706–710 (2009)
38. Li, X., Cai, W., An, J., Kim, S., Nah, J., Yang, D., Piner, R., Velamakanni, A., Jung, I., Tutuc, E., Banerjee, S.K., Colombo, L., Ruoff, R.S.: Large-area synthesis of high-quality and uniform graphene films on copper foils. *Science* **324**, 1312–1314 (2009)
39. Bae, S., Kim, H., Lee, Y., Xu, X., Park, J.-S., Zheng, Y., Balakrishnan, J., Lei, T., Kim, H.R., Song, Y.I., Kim, Y.-J., Kim, K.S., Özyilmaz, B., Ahn, J.-H., Hong, B.H., Iijima, S.: Roll-to-roll production of 30-inch graphene films for transparent electrodes. *Nat. Nanotechnol.* **5**, 574–578 (2010)

Synthesis of Multicationic 1-D Oxide Nanostructures

► Growth of 1-D Oxide Nanostructures

Synthesis of Nanoparticles

► Nanoparticles

► Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods

Synthesis of Subnanometric Metal Nanoparticles

Javier Calvo Fuentes¹, José Rivas² and M. Arturo López-Quintela²

¹Nanogap Sub-Nm-Powder S.A., Milladoiro – Ames (A Coruña), Spain

²Laboratory of Magnetism and Nanotechnology, Institute for Technological Research, University of Santiago de Compostela, Santiago de Compostela, Spain

Synonyms

Atomic cluster; Cluster; Nanocluster; Quantum cluster; Quantum dot

Definition

Metal atomic clusters consist of groups of atoms with well-defined compositions and one or very few stable geometric structures. They represent the most elemental building blocks in nature – after atoms – and are characterized by their size, comparable to the Fermi wavelength of an electron, which makes them a bridge between atoms and nanoparticles or bulk metals, with properties very different from both of them.

Introduction

Typical metal nanoparticles with dimensions from two to several tens of nanometers show smoothly size-dependent properties. However, when particle size becomes comparable to the Fermi wavelength of an electron (~ 0.52 nm for gold and silver), properties of metal clusters are dramatically different from what should be expected if they were due only to their high surface-to-volume ratio. In these subnanometric species, quantum effects are responsible for totally new chemical, optical, and electronic properties such as, for example, magnetism, photoluminescence, or catalytic activity.

The term *nanoparticle* usually refers to any particle of bulk metal with dimensions in the nanoscale. Nanoparticles usually present a core-shell structure with a core of bulk metal surrounded by a shell of disordered atoms, and their enhanced properties are due mainly to their high surface-to-volume ratio. The term *cluster* is used in reference to *subnanometric species* consisting of well-defined structures of metal atoms stabilized by different types of protecting ligands, with sizes below approximately 1–2 nm. In general, clusters can be divided into (1) large clusters, consisting of a core formed by a number of metal atoms in the range $\approx 10\text{--}20$ to $100\text{--}200$ and a protecting shell of strong ligands such as phosphines or thiols; and (2) small clusters formed by a reduced number of atoms (≈ 2 to $10\text{--}20$), which do not need any strong stabilizing ligand and have almost all their atoms on the surface. Due to quantum effects, both kinds of

clusters – large and small – present discrete energy levels and an increasing bandgap with decreasing size (Fig. 1).

Because of this splitting of energies at the Fermi level, clusters show different properties to those of nanoparticles. Cluster properties are highly dependent on their sizes. As an example, the typical surface plasmon band observed in Ag-metal nanoparticles (Fig. 2a) disappears for clusters below approximately 1–2 nm (Fig. 2b, c corresponding to large and small clusters, respectively), indicating that all the conducting electrons are now *frozen* and the *metal* silver loses its typical *metallic character*. At the same time, a clear difference between large and small clusters can be observed: large clusters show a continuous decrease of the absorption band with some small bumps, similar to the absorption displayed by semiconductors (Fig. 2b), and small clusters display well-defined absorption bands indicating a *molecular-like* behavior (Fig. 2c).

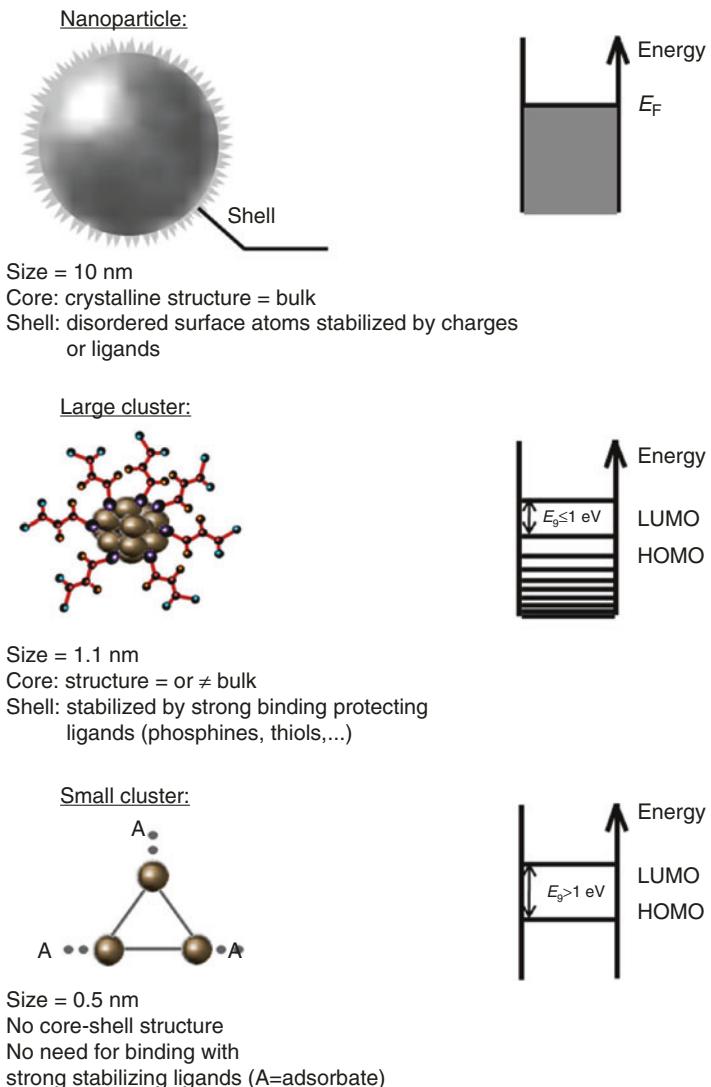
The presence of such size-dependent properties can be well represented on a 3D periodic table of elements – schematically depicted in Fig. 3 – where, between the atomic and the bulk state of every element, a whole range of new materials with their different size-dependent properties appear.

Cluster Structure

It is very difficult to do precise calculations of the electronic structure of metal clusters, particularly when they are larger than just a small number of atoms, and they are in solution stabilized by some capping molecules. Such difficulty can be avoided using simple models, such as the Jellium model [1] firstly developed for clusters in the gas phase. In this model, the real cluster is replaced by an electronic shell structure consisting of a uniform, positively charged *jellium* sphere surrounded by valence electrons. The electrons are considered to be like moving in a mean-field potential occupying energy levels according to the Aufbau principle ($1S^2 | 1P^6 | 1D^{10} | 2S^2 1F^{14} | 2P^6 1G^{18} | \dots$) as represented in Fig. 4. This model gives a

**Synthesis of
Subnanometric Metal
Nanoparticles,**

Fig. 1 Schematic properties of nanoparticles and subnanometric clusters

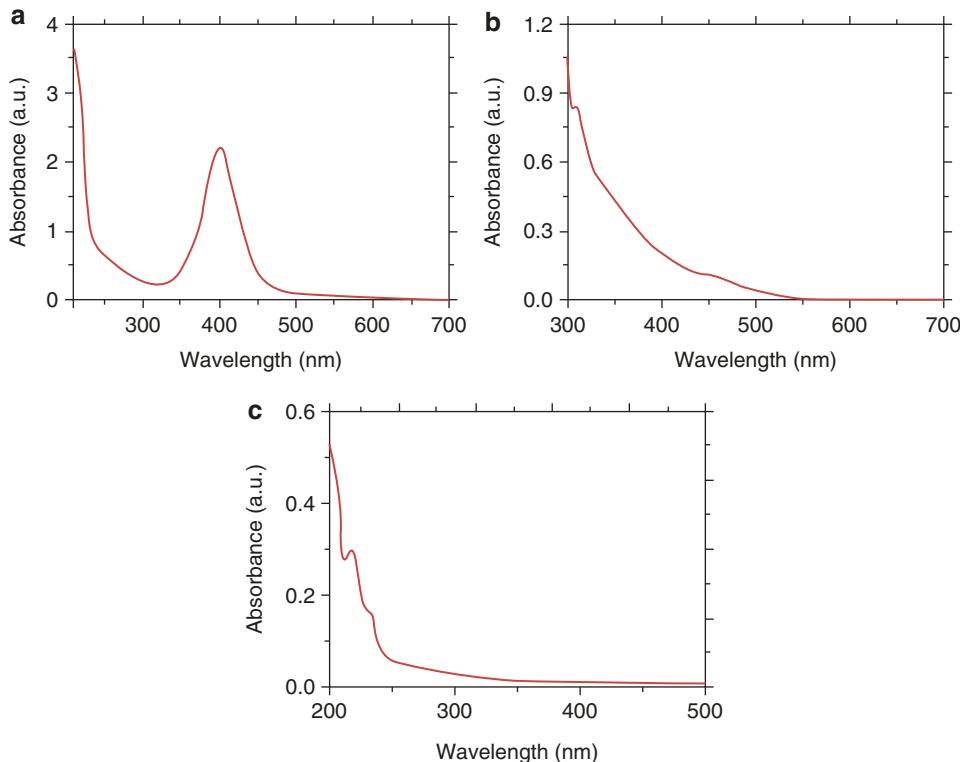


considerably good approximation, preserving many of the physicochemical characteristics of clusters.

The total energy as a function of cluster size, calculated by this approximation, has discontinuities corresponding to electronic shell closures that are consistent with the positions of the peaks appearing in the mass spectrum of the vaporized metal. These peaks are associated with the abundance of exceptionally stable clusters with certain number of atoms: 8, 18, 20, 34, 40, 58, 92, 138, etc., usually referred to as *magic numbers*.

As it was mentioned above, one simple consequence of this quantum-size regime with the presence of discrete states in metal clusters is the appearance of a sizable HOMO-LUMO bandgap similar to that of semiconductors. Such semiconductor-like behavior is particularly significant for smaller clusters (i.e., those with a low number of atoms) with bandgaps widely exceeding 1 eV, as it is shown in Fig. 5 for some selected Ag and Cu clusters.

Some efforts have been made in the last years to find prediction models for ligand-protected



Synthesis of Subnanometric Metal Nanoparticles,
Fig. 2 Absorption spectra of (a) silver nanoparticles displaying the typical plasmon band at ≈ 400 nm, (b) surfactant-protected large silver clusters showing a

typical continuum decrease of the absorption, and (c) strong-ligand-free small silver clusters with well-defined molecular-like absorption bands

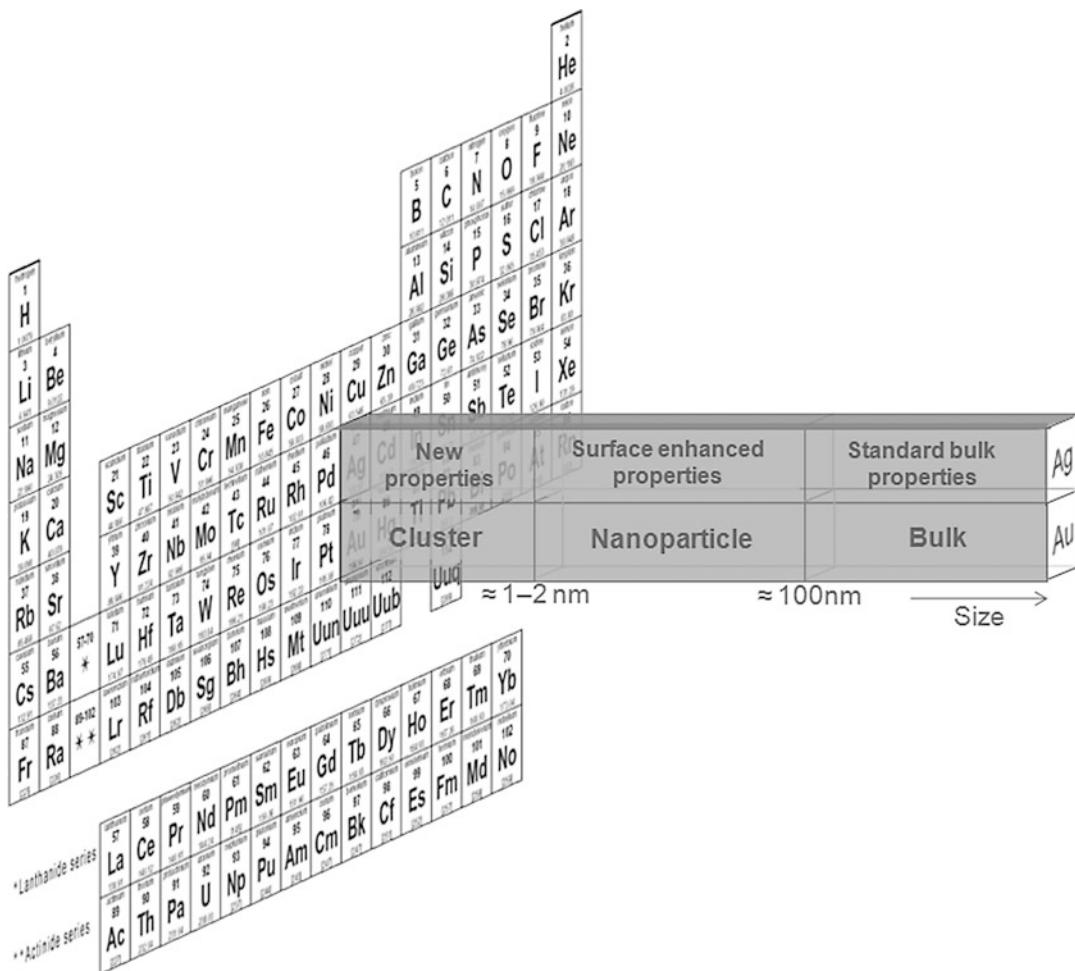
gold-cluster species in solution. In such models, clusters are assumed to be electronically stabilized by ligands that either withdraw electrons from the metal core or are attached as weak Lewis base ligands coordinated to the core surface by dative bonds. In this case, the requirement for an electronically closed shell [3], formulated as $(L_S \cdot A_N X_M)^z$, is

$$n^* = Nv_A - M - z \quad (1)$$

where n^* represents the number of electrons for shell closing of the metallic core, which has to match one of the magic numbers corresponding to strong electron shell closures in an anharmonic mean-field potential, giving rise to stable clusters; N stands for the number of core metal atoms (A); v_A is the atomic valence; M is the number of electron-localizing (or electron-withdrawing)

ligands X , assuming a withdrawal of one electron per ligand molecule; and z represents the overall charge of the complex. The weak ligands represented by L_S may be needed for completing the steric surface protection of the cluster core.

Trying to combine at the same time (1) the mentioned magic numbers required for electronic stabilization of the cluster, (2) the requirements of the Eq. 1, and (3) the fact that solution-phase clusters require a sterically complete protective ligand shell compatible with a compact atomic shell structure for the metallic core complicates enormously the calculations of the structures satisfying such conditions. Furthermore, in cases such as some gold and silver thiolate clusters, the identity of the actual X groups is not clear due to the undefined nature of the surface chemical bonds. There have been only relatively few studies so far to resolve the cluster structures by ab



Synthesis of Subnanometric Metal Nanoparticles, Fig. 3 Size-dependent 3D periodic table of elements

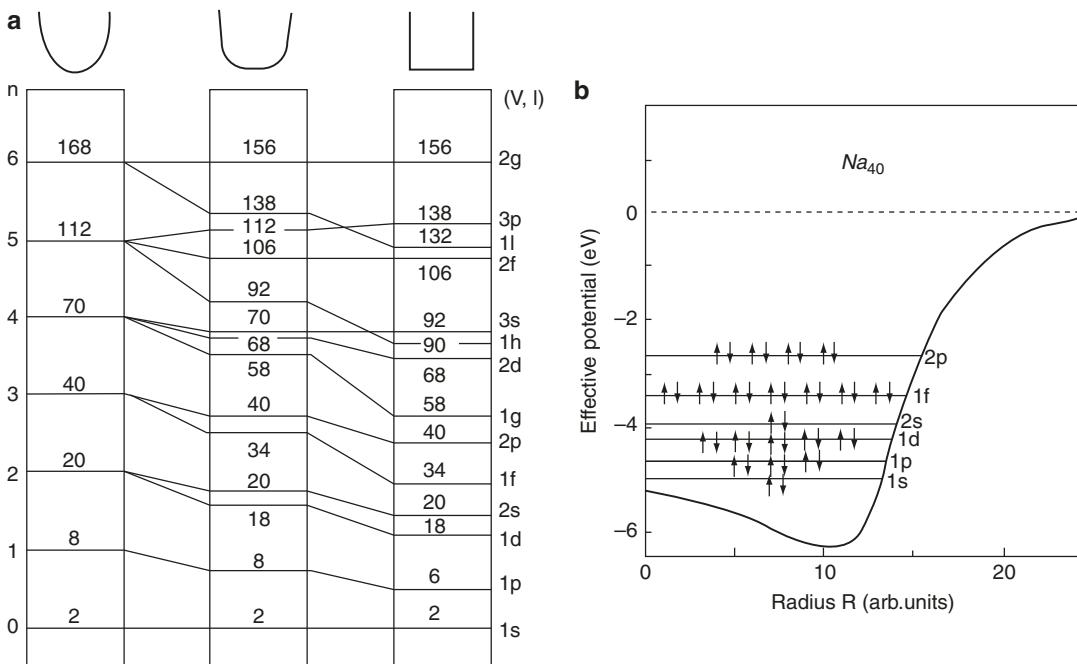
initio calculations. Very recently, Häkkinen's group managed to determine, through large-scale density functional theory (DFT) calculations, [3] the electronic structure of $\text{Au}_{102}(p\text{-MBA})_{44}$ (*p*-MBA, *para*-mercaptobenzoic acid, $\text{SC}_7\text{O}_2\text{H}_5$), an all-thiolate-protected cluster formed by 102 gold atoms, as well as some smaller phosphine-halide- and phosphine-thiolate-protected clusters, as it is shown in Fig. 6.

In all cases, the characteristic bandgap of the clusters is clearly observed. In Table 1, it is shown how this bandgap changes for different cluster sizes and ligands. For comparison purposes, the bandgap calculated from the simple spherical Jellium model is included (for clusters with sizes ≥ 25 atoms, a correction of -0.4 eV was used

because of the anharmonicity observed in large clusters) [4]. It can be observed that, in general, bandgap increases when the cluster size decreases, being a nice agreement between the DFT-calculated bandgap values and those predicted by the simple Jellium model, showing at the same time that ligands play a relatively minor role.

Properties of Metal Clusters

Among the properties that make metal clusters unique and useful in many applications, the following ones can be highlighted here as examples:



Synthesis of Subnanometric Metal Nanoparticles,

Fig. 4 (a) Energy-level occupations for spherical three-dimensional, harmonic, intermediate, and square-well potentials. (b) Self-consistent effective potential of a

Jellium sphere corresponding to Na_{40} with the electron occupation of the energy levels (Reprinted with permission from Ref. [2]. Copyright 1993 by the American Physical Society)

Photoluminescence

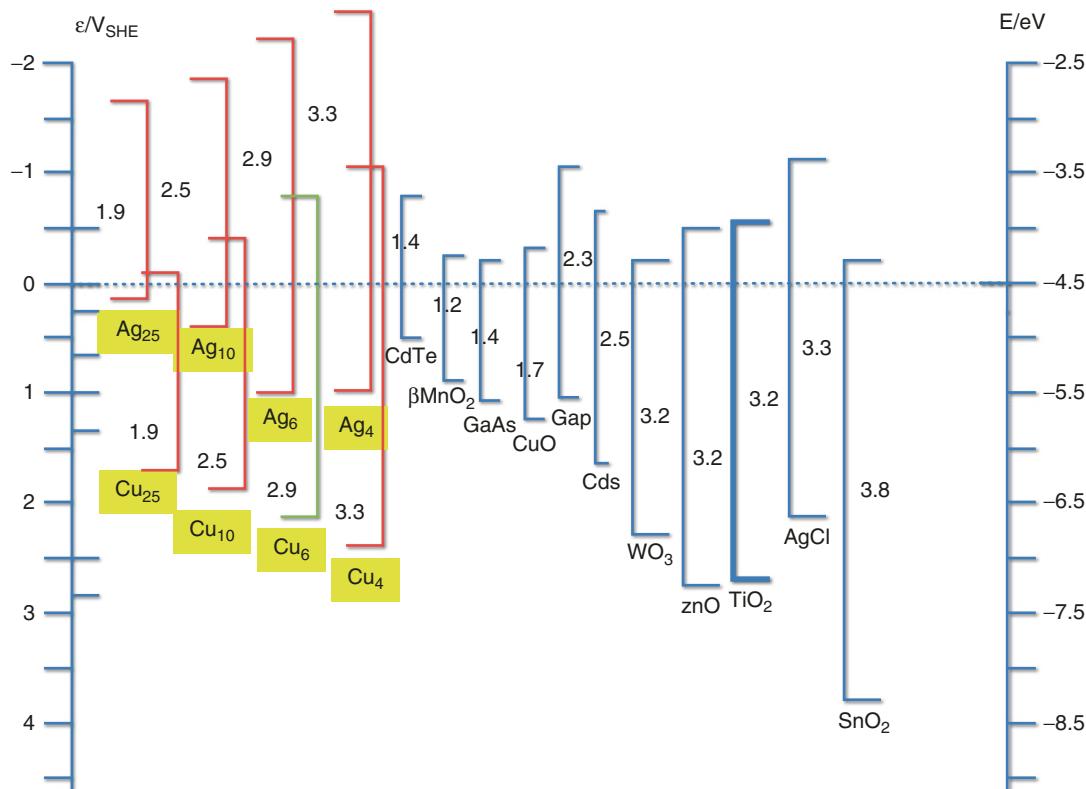
The molecular-like behavior of few-atom noble metal clusters, and the mentioned bandgap similar to that of semiconductors, allows the possibility of size-tunable electronic intraband transitions [5].

The appearance of these optical transitions with energies ranging from the UV-visible to the infrared region makes noble metal quantum dots ideal potential candidates for applications such as fluorescent labeling in biology or light-emitting sources in nanoelectronics. As an example, Fig. 7 shows the fluorescence of aqueous solutions of Cu_N clusters ($N < 14$) excited at 296 nm with quantum yields over 10 %, which are stable for more than 2 years.

Even though, in cases where the quantum yield is lower than that cited in previous example, fluorescent noble metal clusters are a very interesting alternative to semiconductor quantum dots as biological labels, due not only to their small hydrodynamic size and inert nature but especially because of their biocompatibility and also because

they present reduced photobleaching compared to organic fluorophores. Water-soluble gold clusters have been, therefore, explored as fluorescent labels in biological experiments. Within this approach, Parak's group described [7] the use of stable gold clusters capped with dihydrolipoic acid (DHLA) and conjugated through EDC chemistry to biologically relevant molecules such as PEG, BSA, avidin, or streptavidin for biolabeling. Unconjugated clusters can also be nonspecifically taken up by living cells (Fig. 8).

The last contributions in this field exploit a large Stokes shift displayed by a new type of specially synthesized clusters which are particularly interesting for biosensing applications since they present an intense fluorescent emission in the red spectra (highly desired due to its large signal-to-noise ratio). While the origin of this large Stokes shift – usually associated to a metal-ligand charge transfer – for this new class of materials is not totally understood, the technical development is going fast due to the facile synthesis procedures



Synthesis of Subnanometric Metal Nanoparticles,

Fig. 5 Schematic comparison between bandgaps of some silver and copper clusters (M_N , N = number of atoms) and those of well-known semiconductors.

already developed [8, 9]. For example, Biang et al. [8] prepared histidine (His)- and 11-mercaptopoundecanoic acid (MUA)-capped gold clusters, ($Au_{17}MUAH_4Hiss_{22}$), with strong fluorescence around 600 nm and a large lifetime $\sim 7\mu s$ (see Fig. 9). The as prepared clusters were tested in HeLa cell lines and a marginal toxicity was observed. Fluorescence microscopy showed that such clusters were internalized into the cell cytoplasm by endocytosis. These novel clusters are ideal for deep tissue penetration and biological imaging by fluorescence microscopy. Kong et al. [9] reported another interesting example of highly fluorescent Au clusters using bovine pancreatic ribonuclease A (RNase A) as bio-template. The RNase A-encapsulated Au clusters show a strong fluorescence (quantum yield of 12 %) in the near-infrared region (682 nm) and exhibit a large Stokes shift (210 nm) and a single dominant

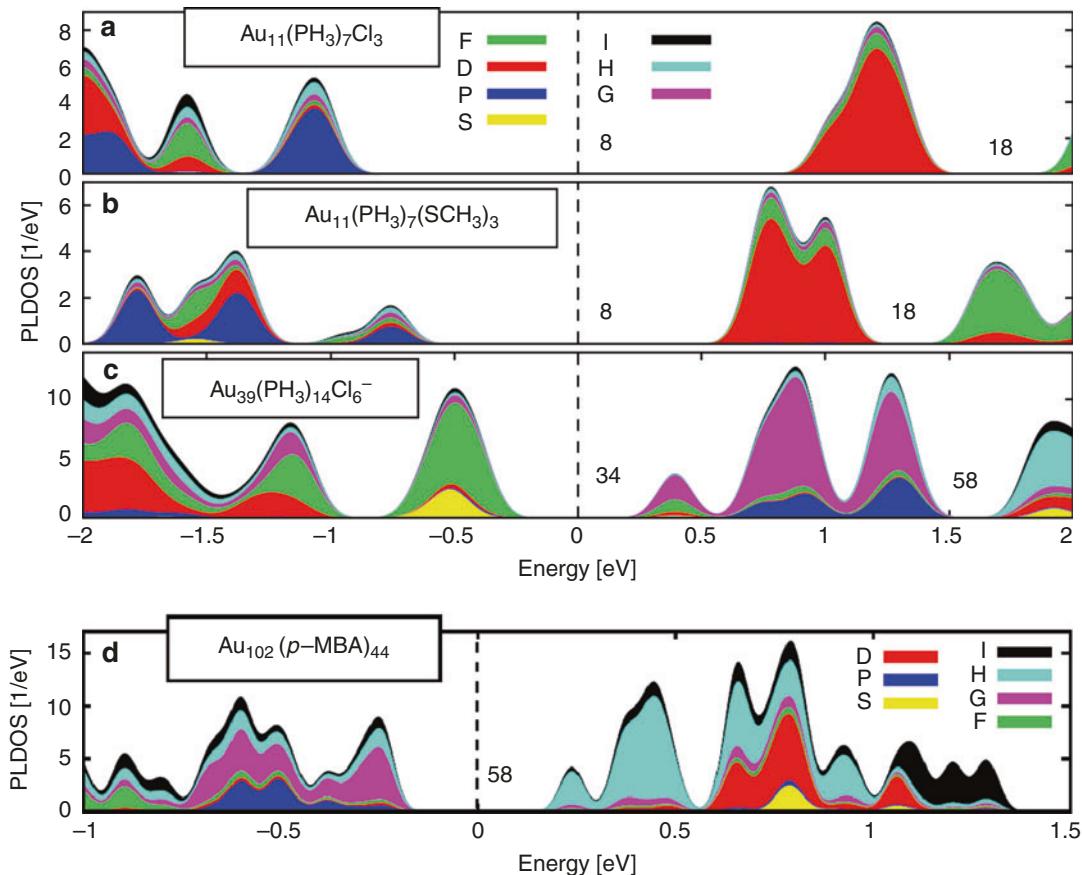
Bandgaps (E_g) were calculated from the spherical Jellium model ($E_g = E_F/N^{1/3}$; E_F = Fermi level), and the position of the conduction band, E_{CB} , was estimated by the formula $E_{CB} = \chi - E_F - \frac{1}{2} E_g$ (χ = electronegativity)

fluorescence lifetime of $\sim 1.5\mu s$. By coupling these Au clusters with vitamin B₁₂, a multifunctional platform was developed, suitable for simultaneous targeting and imaging of cancer at the cellular level using Caco-2 cell lines as in vitro model.

S

Catalysis

Small clusters have been found to own a catalytic activity not observed in their bulk analogs or even nanoparticles, which makes them very attractive as new catalytic materials. Quantum chemical calculations indicate that such high reactivity is due to undercoordination of the metal atoms forming the cluster [10]. Several families of metal clusters have been found to possess surprisingly high and selective catalytic activity when immobilized on a support. For example, platinum clusters with 8–10 atoms can be used as catalysts for the oxidative dehydrogenation of



Synthesis of Subnanometric Metal Nanoparticles.
Fig. 6 Electronic structure analysis of gold clusters. The angular-momentum-projected local electron density of states (PLDOS) (projection up to the l symmetry, i.e., $l = 6$) for the gold core in $\text{Au}_{11}(\text{PH}_3)_7\text{Cl}_3$ (**a**), $\text{Au}_{11}(\text{PH}_3)_7(\text{SCH}_3)_3$ (**b**), $\text{Au}_{39}(\text{PH}_3)_{14}\text{Cl}_6^-$ (**c**), and $\text{Au}_{102}(p\text{-MBA})_{44}$ (**d**). The zero energy corresponds to the

middle of the HOMO-LUMO gap (dashed line). For plotting, each individual electron state is displayed by a Gaussian smoothing of 0.07 eV (0.03 eV in **d**). Shell-closing number is indicated for each case (For more details see Ref. [3]. Copyright 2008 National Academy of Sciences, USA)

Synthesis of Subnanometric Metal Nanoparticles, Table 1 Comparison between: (1) experimentally determined bandgaps for free gas phase gold cluster anions from photoelectron spectroscopy; (2) theoretical (DFT) values for HOMO-LUMO gaps of passivated gold cluster compounds that correspond to 8, 34, and 58 conduction-electron shell closings; and (3) bandgaps calculated through the Jellium model for clusters with the same number of gold atoms

	Experiment		Theory		Jellium model
Shell closing	Cluster	Gap (eV)	Cluster compound	Gap (eV)	$E_g = 5.32/N^{1/3}$ (eV)
8e ($1\text{S}^21\text{P}^0$)			$\text{Au}_{11}(\text{PH}_3)_7(\text{SMe})_3$	1.5	2.4
8e			$\text{Au}_{11}(\text{PH}_3)_7\text{Cl}_3$	2.1	2.4
8e			$\text{Au}_{13}(\text{PH}_3)_{10}\text{Cl}_2^{3+}$	1.8	2.3
8e			$\text{Au}_{25}(\text{SMe})_{18}^-$	1.2	1.3 ^a
34e ($8\text{e} + 1\text{D}^{10}2\text{S}^21\text{F}^{14}$)	Au_{34}^-	1.0	$\text{Au}_{39}\text{Cl}_6(\text{PH}_3)_{14}^-$	0.8	1.0 ^a
58e ($34\text{e} + 2\text{P}^61\text{G}^{18}$)	Au_{58}^-	0.6	$\text{Au}_{102}(\text{p-MBA})_{44}$	0.5	0.6 ^a
58e			$\text{Au}_{102}(\text{SMe})_{44}$	0.5	0.6 ^a

Adapted from Ref. [3]. Copyright 2008 National Academy of Sciences, USA

^aFor these clusters, a correction of -0.4 eV was used – see text

propane [11], while gold clusters with 6–10 atoms have been shown to be highly active for catalyzing propene epoxidation [12]. Recently, Harding et al. [13] studied the control and tunability of the catalytic oxidation of CO by Au_{20} clusters deposited on MgO surfaces, finding that the active site



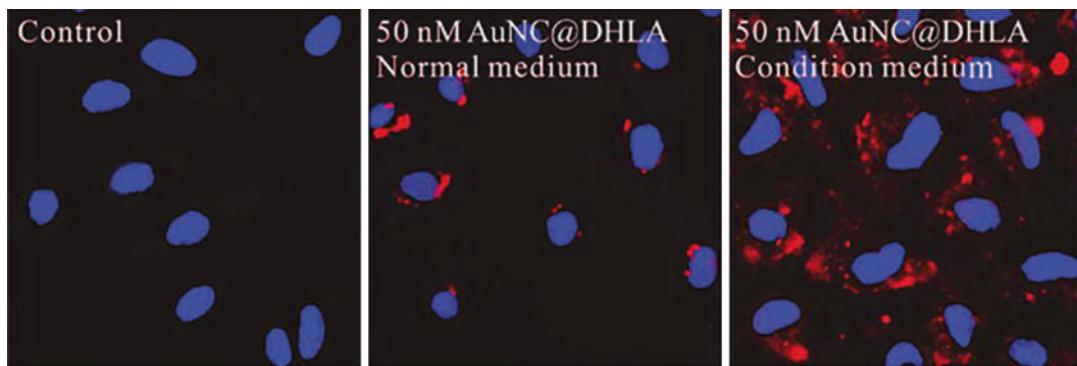
Synthesis of Subnanometric Metal Nanoparticles,

Fig. 7 Blue emission observed for an aqueous solution of copper clusters ($N \leq 14$ atoms) under an excitation wavelength of 296 nm (Reprinted with permission from Ref. [6]. Copyright 2010 American Chemical Society)

on the cluster is characterized by enhanced electron density, which activates the adsorbed O_2 molecule and promotes the bonding of CO.

Small metal clusters have been also found to display electrocatalytic activities different from the material in bulk or as nanoparticles [14]. Although this is a field not too much explored so far, these electrocatalytic properties make metal clusters to be promising materials in fuel-cell applications. As another interesting example of such properties, a recent report of their use to prevent pathologies, such as fetal alcohol syndrome, [15] is highlighted here: It has been proved that some clusters are able to electrocatalyze the oxidation of alcohols and prevent its cytotoxicity under physiological conditions and at very low potentials, like those found on living cells, as it is schematically represented in Fig. 10.

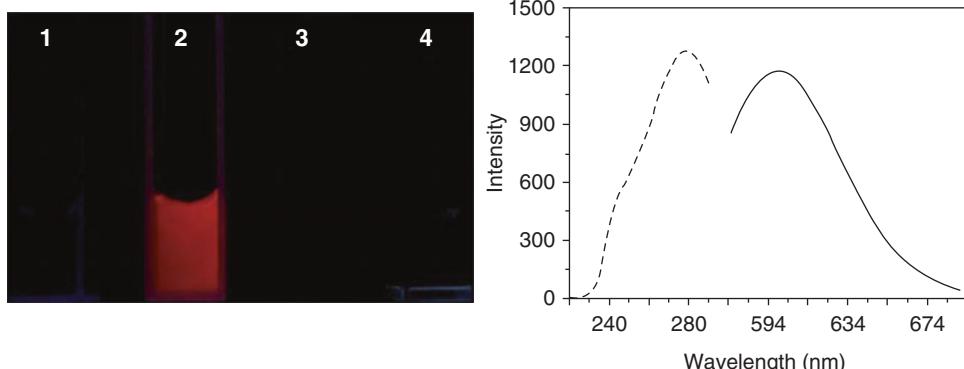
Differently from NPs and bulk surfaces, clusters present an optimum window for their size-dependent catalytic activity which is different for each type of reaction. Recently, Corma et al. carried on a deep theoretical and experimental study to unravel the exceptional catalytic activity shown by small Au clusters (between 5 and 10 atoms) for the aerobic thiophenol oxidation, comparable to sulfhydryl oxidase enzymes [16]. Under this approach, the existence of this optimum window size is qualitatively explained in terms of the position of the HOMO-LUMO bands of gold clusters and the bonding to the



Synthesis of Subnanometric Metal Nanoparticles,

Fig. 8 Nonspecific uptake of unconjugated fluorescent DHLA-capped Au clusters by human endothelial cells. Cell nuclei were stained to yield blue fluorescence. Red

fluorescence corresponds to the clusters. (see Ref. [7] for more information. Copyright 2009 American Chemical Society. Reprinted with permission)



Synthesis of Subnanometric Metal Nanoparticles,
Fig. 9 A large Stokes shift displayed by specifically synthesized clusters allows to obtain intense fluorescent emission in the *red* spectra which, due to its high signal-to-noise

ratio, is highly desired for biosensing applications (see Ref. [8] for a complete description. Reprinted with permission of The Royal Society of Chemistry)

thiols. A weak bonding to thiols is necessary for the catalysis to take place, thus activating the thiol, but a strong bonding inhibits the reaction. While big clusters –with a low HOMO band- and nanoparticles strongly bond to thiols, very small clusters (<5 atoms) –with a high HOMO band– cannot bond to them. This leads to the observed fact that only medium-size clusters, between 5 and 10 atoms, have the weak bonding necessary to activate the thiol.

Synthesis of Metal Clusters

As it also occurs in the case of nanoparticles, there are two main approaches to the synthesis of metal clusters: top-down and bottom-up.

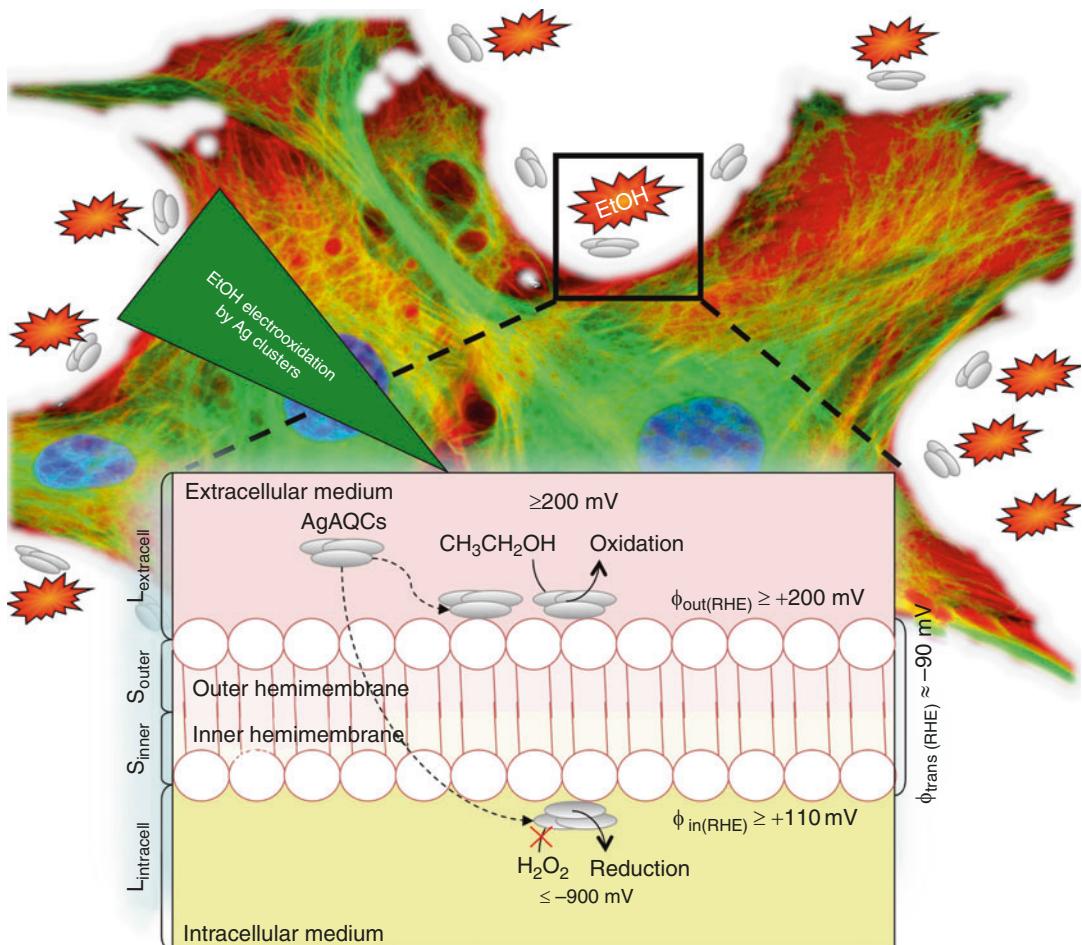
Bottom-Up

In bottom-up approaches, clusters are built from smaller structures at atomic level such as single atoms or ions (see Fig. 11). In this case, the most commonly used techniques differ depending on the size of the clusters to be synthesized. Wet chemical preparation of clusters includes the chemical reduction of metal salts in the presence of strong binding ligands [4, 17, 18] or cages [19] and by kinetic control using, e.g., microemulsions [20–22], electrochemical methods [6, 14, 23], etc.

For the synthesis of large clusters, a large variety of strong, protecting ligands can be used to

control the growth of the primarily formed clusters and to stabilize them. Within this approach, we can mention the size-controlled synthesis of glutathione-capped gold clusters [24, 25]. The key of this beautiful design is based on the use of a reducing agent, gaseous CO, which establishes a mild reaction environment for a slow and well-controlled growth of Au clusters. The fine-tune kinetics is achieved by precisely controlling the pH, leading to the formation of a monodisperse set of products: $\text{Au}_{10-12}\text{GSH}_{10-12}$, $\text{Au}_{15}\text{GSH}_{13}$, $\text{Au}_{18}\text{GSH}_{14}$, and $\text{Au}_{25}\text{GSH}_{18}$ (Fig. 12).

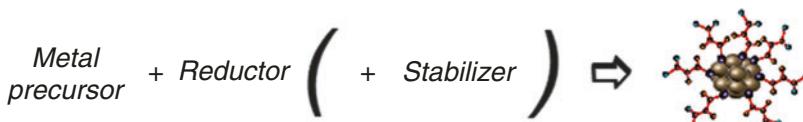
For small clusters, the kinetic control, based on slow reaction rates provided either by soft reducing agents or low concentrations (as e.g., those provided by microemulsions, electrochemical methods, etc.), is the key point to their synthesis. It is generally believed that the synthesis of small clusters is extremely difficult to achieve because of the highly precise control of experimental conditions required to stop and then isolate the clusters as soon as they are formed. That belief is based on the theory of nucleation and growth (NGT), according to which, nuclei formed during the first steps of the chemical synthesis are only stable when they grow over a specific size called critical nucleus. Below that size, nuclei dissolve because of their large Laplace pressure. Above that size, they continue growing in order to reduce their surface energy by different mechanisms like autocatalysis, Ostwald ripening, etc., until the growth



Synthesis of Subnanometric Metal Nanoparticles,

Fig. 10 Schematic representation of the electrocatalysis of the oxidation of ethanol by silver clusters on the cellular

membrane, preventing the alcohol toxicity in living cells (For more details, see Ref. [15]. Copyright 2010 American Chemical Society. Reprinted with permission)



Synthesis of Subnanometric Metal Nanoparticles, Fig. 11 Schematic representation of bottom-up synthesis

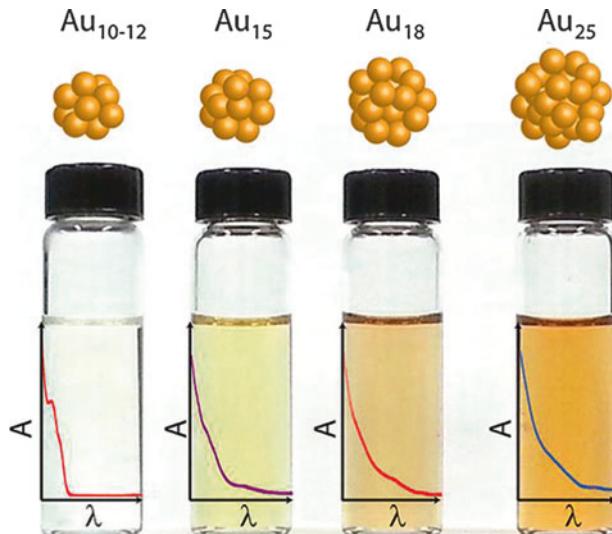
is stopped by capping agents or other templates. However, such arguments, which correctly can be applied to fast reaction rates or high-temperature synthesis of nanoparticles and larger clusters having the same structure of the bulk material, are not correct when they apply to very small clusters, as those produced by kinetic control. As it was

previously mentioned, numerous theoretical and experimental reports indicate that clusters can be especially stable because of their particular electronic and geometric structures different from the bulk.

In such a case, the use of macroscopic thermodynamic arguments, like those used in the theory

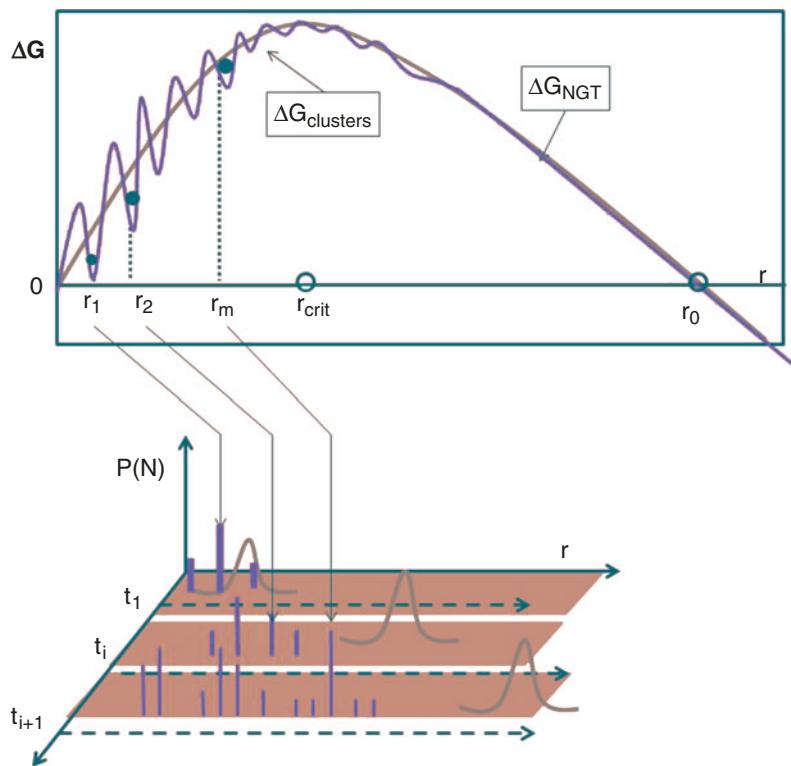
Synthesis of Subnanometric Metal Nanoparticles,

Fig. 12 Glutathione-capped gold clusters, synthesized with a simple procedure (Reprinted from Ref. [25] with permission. Copyright 2014 American Chemical Society)



Synthesis of Subnanometric Metal Nanoparticles,

Fig. 13 Schematic representation of the variation of free energy during the chemical synthesis of metal nanoparticles following the classical NGT compared to the one modified including the cluster effect in the NGT. Size distribution for cluster formation becomes a discrete magnitude that reflects the large stability of magic number clusters (Reprinted from Ref. [26] with permission from Elsevier)



of nucleation and growth, cannot be applied, as it was recently published in a revision of the NGT that takes into account the existence of stable clusters [26]. Figure 13 shows a comparison between the free energy evolution throughout a reaction of NP formation in solution, obtained using the

classical NGT, and the modified one including clusters' existence. The cost in free energy for the growth of clusters evolves through a stepwise stability wells which correspond to the enhanced stability of stable clusters with magic numbers. Under this approach, clusters do not obey the

rule, assumed in the classical NGTs, of being continuously formed and dissolved and correspondingly the size distribution, $P(N)$, which for classical NGT evolves in time as a continuous function of size becomes for the clusters' formation a discrete magnitude that reproduces the large stability of magic number clusters.

As said before, in a typical synthesis of nanoparticles, a fast reduction reaction is used, which means that the driving force (which can be represented by the difference in the electrochemical potentials of the metal to be synthesized and the reducing agent) is enough to drive the system directly to the first stable nucleus (r_{crit}) having the same structure of the bulk. However, when a small reaction rate is used, stable clusters like r_1 , r_2 , etc., are produced, and with an adequate control of the reaction kinetics, it is possible to stop the reaction at different times and, in this way, to obtain stable metal clusters with the desired size.

Following this scheme, silver clusters with less than 10 atoms have been synthesized, for example, in water-in-oil microemulsions consisting of a mixture of sodium bis(2-ethylhexyl) sulfosuccinate, isoctane, and water [21]. Synthesis conditions necessary for the appearance and isolation of stable clusters were achieved using small concentrations and a mild reduction agent as sodium hypophosphite monohydrate. Silver clusters produced in this way were found not only to be stable for years, but also to have a large bandgap (≈ 2.3 eV) and to present molecular-like paramagnetic properties and an intense photoluminescence.

In the same way, using electrochemical techniques, very small clusters can be synthesized, providing a good control of the cluster size. For example, stable poly(*N*-vinylpyrrolidone), PVP-protected gold clusters with only 2 or 3 atoms, could be synthesized through a simple electrochemical method based on the anodic dissolution of a gold electrode in the presence of the homopolymer PVP and subsequent electroreduction of the Au-PVP complexes [23]. The resulting clusters (Au_2 and Au_3) are the smallest ones that can be prepared and they display fluorescent (Au_2) and paramagnetic (Au_3) properties. By a similar procedure, highly

fluorescent copper clusters stabilized with tetrabutylammonium nitrate, with less than 14 atoms, have also been reported [6]. These clusters can be dispersed in different solvents (both polar and apolar), are stable for several years, and display also similar fluorescent properties than those synthesized in microemulsions [22].

Finally, it is worth mentioning the use of hard cages to inhibit the growth and synthesize small clusters, as it was demonstrated very recently by Royon et al. [19], showing that silver-containing glasses can be used for preparing fluorescent materials, reducing the Ag ions in the matrix with laser pulses. In this way, they were able to prepare perennial high-capacity 3D-optical recording media.

Top-Down

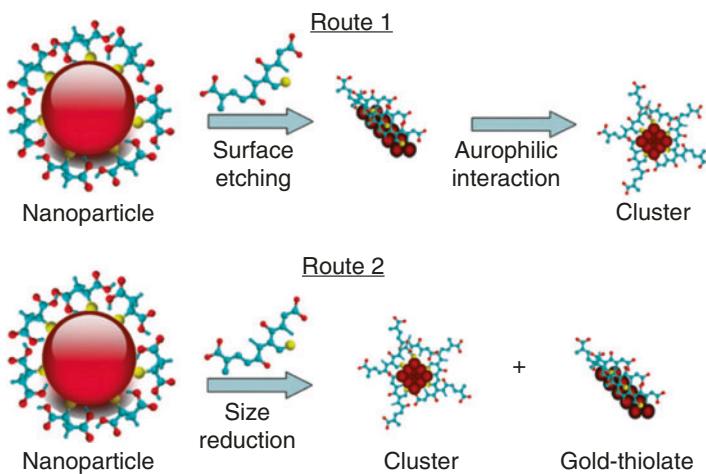
In top-down approaches, clusters are synthesized from larger precursors such as nanoparticles or bulk metal. The most common top-down technique used for the synthesis of clusters is etching of nanoparticles (Fig. 14). In this technique, clusters are synthesized using the etching capacity of some ligands (like thiols), which remove the surface atoms of metallic nanoparticles or break them into smaller pieces leading to stable quantum clusters characterized by shell-closing magic numbers. Figure 13 can also help to clarify and understand this important top-down process in energy terms. Nanoparticles with sizes close to the critical nucleus ($r_{\text{crit}} < r < r_0$) can be etched, yielding stable clusters whose potential wells are located below the DG of the etched NPs ($DG > 0$, i.e., for sizes in the metastable region), making the process energetically favorable. Experimentally, etching is widely used as a top-down technique.

For example, etching of mercaptosuccinic acid-protected gold nanoparticles with excess glutathione has been used to yield small photoluminescent gold clusters [27]. In such work, clusters, either with 8 or 25 gold atoms, were produced by etching of previously synthesized nanoparticles with 4–5 nm core diameter. Selection of the final cluster size was possible by only varying the etching pH from ~ 3 (for 25 atom clusters) to 7–8 (for 8 atom clusters).

Multivalent coordinating polymers such as polyethylenimine can also be used to etch

Synthesis of Subnanometric Metal Nanoparticles

Fig. 14 Schematic representation of two possible routes for the formation of gold clusters by etching of mercaptosuccinic acid-capped gold nanoparticles (Reprinted from Ref. [27] with kind permission from Springer Science + Business Media B.V)



preformed colloidal gold nanocrystals, producing highly fluorescent water-soluble nanoclusters formed by 8 gold atoms [28].

In summary, it can be said that by appropriate use of the typical techniques developed for the synthesis of nanoparticles, clusters of different sizes down to 2 atoms can be produced, isolated, and used as stable materials with novel properties, which differ very strongly from the bulk materials and nanoparticles due to their totally different electronic and geometrical structure.

Cross-References

- ▶ Dry Etching Processes
- ▶ Nanoparticles
- ▶ Optical and Electronic Properties

References

- Chou, M.Y., Cleland, A., Cohen, M.L.: Total energies, abundances, and electronic shell structure of lithium, sodium, and potassium clusters. *Solid State Commun.* **52**, 645–648 (1984)
- de Heer, W.A.: The physics of simple metal clusters: experimental aspects and simple models. *Rev. Mod. Phys.* **65**, 611–676 (1993)
- Walter, M., Akola, J., Lopez-Acevedo, O., Jadzinsky, P.D., Calero, G., Ackerson, C.J., Whetten, R.L., Grönbeck, H., Häkkinen, H.: A unified view of ligand-protected gold clusters as superatom complexes. *Proc. Natl. Acad. Sci.* **105**, 9157–9162 (2008)
- Zheng, J., Zhang, C., Dickson, R.M.: Highly fluorescent, water-soluble, size-tunable gold quantum dots. *Phys. Rev. Lett.* **93**, 077402 (2004)
- Zheng, J., Nikovich, P.R., Dickson, R.M.: Highly fluorescent noble metal quantum dots. *Annu. Rev. Phys. Chem.* **58**, 409–431 (2007)
- Vilar-Vidal, N., Blanco, M.C., López-Quintela, M.A., Rivas, J., Serra, C.: Electrochemical synthesis of very stable photoluminescent copper clusters. *J. Phys. Chem. C* **114**, 15924–15930 (2010)
- Lin, C.-A.J., Yang, T.-Y., Lee, C.-H., Huang, S.H., Sperling, R.A., Zanella, M., Li, J.K., Shen, J.-L., Wang, H.-H., Yeh, H.-I., Parak, W.J., Chang, W.H.: Synthesis, characterization, and bioconjugation of fluorescent gold nanoclusters toward biological labeling applications. *ACS Nano* **3**, 395–401 (2009)
- Biang, P., Zhou, P., Liu, Y., Ma, Z.: One-Step fabrication of intense red fluorescent gold nanoclusters and their application in cancer cell imaging. *Nanoscale* **5**, 6161–6166 (2013)
- Kong, Y., Chen, J., Gao, F., Brydson, R., Johnson, B., Heth, G., Zhang, Y., Wu, L., Zhou, D.: Near-infrared fluorescent ribonuclease-A-encapsulated gold nanoclusters. *Nanoscale* **5**, 1009–10017 (2013)
- Heiz, U., Landman, U.: *Nanocatalysis (Nanoscience and Technology)*. Springer, Berlin (2007)
- Vajda, S., Pellin, M.J., Greeley, J.P., Marshall, C.L., Curtiss, L.A., Ballentine, G.A., Elam, J.W., Catillon-Mucherle, S., Redfern, P.C., Mahmood, F., Zapal, P.: Subnanometre platinum clusters as highly active and selective catalysts for the oxidative dehydrogenation of propane. *Nat. Mater.* **8**, 213–216 (2009)
- Lee, S., Molina, L.M., López, M.J., Alonso, J.A., Hammer, B., Lee, B., Seifert, S., Winans, R.E., Elam, J.W., Pellin, M.J., Vajda, S.: Selective propene epoxidation on immobilized Au_{6–10}clusters: the effect of hydrogen and water on activity and selectivity. *Angew. Chem. Int. Ed.* **48**, 1467–1471 (2009)
- Harding, C., Habibpour, V., Kunz, S., Farnbacher, A. N.-S., Heiz, U., Yoon, B., Landman, U.: Control and

- manipulation of gold nanocatalysis: effects of metal oxide support thickness and composition. *J. Am. Chem. Soc.* **131**, 538–548 (2009)
14. Rodríguez-Vázquez, M.J., Blanco, M.C., Lourido, R., Vázquez-Vázquez, C., Pastor, E., Planes, G.A., Rivas, J., López-Quintela, M.A.: Synthesis of atomic gold clusters with strong electrocatalytic activities. *Langmuir* **24**, 12690–12694 (2008)
 15. Selva, J., Martínez, S.E., Buceta, D., Rodríguez-Vázquez, M.J., Blanco, M.C., López-Quintela, M.A., Egea, G.: Silver sub-nanoclusters electrocatalyze ethanol oxidation and provide protection against ethanol toxicity in cultured mammalian cells. *J. Am. Chem. Soc.* **132**, 6947–6954 (2010)
 16. Corma, A., Concepción, P., Boronat, M., Sabater, M. J., Navas, J., Yacamán, M.J., Larios, E., Posadas, A., López-Quintela, M.A., Buceta, D., Mendoza, E., Guilera, G., Mayoral, A.: Exceptional oxidation activity with size-controlled supported gold clusters of low atomicity. *Nat. Chem.* **5**, 775–781 (2013)
 17. Schaeffer, N., Tan, B., Dickinson, C., Rosseinsky, M.J., Laromaine, A., McComb, D.W., Stevens, M.M., Wang, Y., Petit, L., Barentin, C., Spiller, D.G., Cooper, A.I., Lévy, R.: Fluorescent or not? Size-dependent fluorescence switching for polymer-stabilized gold clusters in the 1.1–1.7 nm size range. *Chem. Commun.* **34**, 3986–3988 (2008)
 18. Negishi, Y., Nobusada, K., Tsukuda, T.: Glutathione-protected gold clusters revisited: bridging the gap between gold(I)-thiolate complexes and thiolate-protected gold nanocrystals. *J. Am. Chem. Soc.* **127**, 5261–5270 (2005)
 19. Royon, A., Bourhis, K., Bellec, M., Papon, G., Bousquet, B., Deshayes, Y., Cardinal, T., Canioni, L.: Silver clusters embedded in glass as a perennial high capacity optical recording medium. *Adv. Mater.* **22**, 5282–5286 (2010)
 20. López-Quintela, M.A.: Synthesis of nanomaterials in microemulsions: formation mechanisms and growth control. *Curr. Opin. Colloid Interface Sci.* **8**, 137–144 (2003)
 21. Ledo-Suárez, A., Rivas, J., Rodríguez-Abreu, C.F., Rodríguez, M.J., Pastor, E., Hernández-Creus, A., Oseroff, S.B., López-Quintela, M.A.: Facile synthesis of stable subnanosized silver clusters in microemulsions. *Angew. Chem. Int. Ed.* **46**, 8823–8827 (2007)
 22. Vázquez-Vázquez, C., Bañobre-López, M., Mitra, A., López-Quintela, M.A., Rivas, J.: Synthesis of small atomic copper clusters in microemulsions. *Langmuir* **25**, 8208–8216 (2009)
 23. Santiago González, B., Rodríguez, M.J., Blanco, C., Rivas, J., López-Quintela, M.A., Martinho, J.M.G.: One step synthesis of the smallest photoluminescent and paramagnetic PVP-protected gold atomic clusters. *Nano Lett.* **10**, 4217–4221 (2010)
 24. Yu, Y., Chen, X., Yao, Q., Yu, Y., Yan, N., Xie, J.: Scalable and Precise Synthesis of Thiolated Au_{10-12} , Au_{15} , Au_{18} , and Au_{25} Nanoclusters via pH controlled CO Reduction. *Chem. Mater.* **25**, 946–952 (2013)
 25. Stamplecoskie, K.G., Kamat, P.V.: Size-dependent excited state behavior of glutathione -capped gold clusters and their light-harvesting capacity. *J. Am. Chem. Soc.* **136**, 11093–11109 (2014)
 26. Piñeiro Redondo, Y., Buceta, D., Huseyinova, S., Cuerva, M., Perez Mariño, A., Dominguez, B., Calvo, J., López-Quintela, M. A.: Large stability and high catalytic activities of sub-nm metal (0) clusters: implications into the nucleation and growth theory. *J. Colloid Interface Sci.* **449**, 279–285 (2015)
 27. Habeeb Muhammed, M.A., Ramesh, S., Sinha, S.S., Pal, S.K., Pradeep, T.: Two distinct fluorescent quantum clusters of gold starting from metallic nanoparticles by pH-dependent ligand etching. *Nano Res.* **1**, 333–340 (2008)
 28. Duan, H., Nie, S.: Etching colloidal gold nanocrystals with hyperbranched and multivalent polymers: a new route to fluorescent and water soluble atomic clusters. *J. Am. Chem. Soc.* **129**, 2412–2413 (2007)

Synthesized Conductance Injection

► Dynamic Clamp

Synthesized Ionic Conductance

► Dynamic Clamp

Synthesized Synaptic Conductance

► Dynamic Clamp

Synthetic Biology

Soichiro Tsuda

School of Chemistry, University of Glasgow,
Glasgow, UK

S

Synonyms

[Synthetic genomics](#)

Definition

Synthetic biology is the design and construction of biological components (e.g., enzymes, gene circuits, and whole cells) from scratch or from standardized parts.

Overview

The term “synthetic biology” was first used in 1980 by Barbara Hobom to describe genetically engineered bacteria using recombinant DNA technology. In 2000, the term reappeared again by Eric Kool and others at the annual meeting of the American Chemical Society to refer to synthesis of unnatural chemical products (e.g., organic molecules) as an extension of synthetic chemistry [1].

Although the scope of synthetic biology has explosively expanded in the last decade, synthetic biology can be best described as “efforts to redesign life,” that is, the design and construction of biological components (e.g., enzymes, gene circuits, and whole cells) using living cells. Similar approaches have already been taken long before the advent of synthetic biology, for example, by genetic engineering, which is one of the roots of synthetic biology. However, what makes synthetic biology distinctive from other preceding biological research fields is its strong engineering perspective. While genetic engineering focused on modifying individual genes and pathways, synthetic biology does on whole systems of genes and gene products. It aims at designing and constructing the behavior of organisms to perform new tasks [2]. Computer engineering analogy is often used to explain the goal of synthetic biology: In order to build a computer, it involves several hierarchical layers. The most fundamental layer of computers is the electrical components, such as resistor, transistor, capacitor, etc. Based on them, a higher-level structure, logic gates (AND, OR, NOT, XOR, etc.) are built, which are further integrated into even upper layer, integrated circuits, and so on. Biological systems can be described similarly: DNA, RNA, and proteins are the most fundamental components of biological cells. These components form gene circuit

networks, which are parts of intracellular signal pathways, and so on. Thus, synthetic biology research involves the developments of biological components in each layer for engineering new biological functions: standardized fundamental biological parts, gene circuit construction, and whole biological cells, and biological and chemical products synthesized by living cells.

Standardized Biological Parts

To engineer complex systems, the standardization of components used in the systems (in this case biological systems) is necessary. Tom Knight and Drew Endy proposed the BioBrick standard biological parts as basic units for synthetic biological systems and set up an online registry of BioBrick parts where anyone can contribute in an open-source manner [3–5]. It provides a collection of well-characterized and standardized building blocks that can be used for the “plug-and-play” design and construction of unnatural biological systems. Using BioBrick standard parts, a biological engineer can program a living cell, as an electrical engineer can program an electrical circuit using electrical components. BioBrick parts are mostly DNA sequences with functionality, such as promoters, plasmids, and ribosome binding sites. Not only natural components, biologically inspired molecules can be BioBrick parts, for example, DNA analog PNA, in which the sugar-phosphate backbone is altered by *N*-(2-aminoethyl)-glycine units linked by peptide bonds [6].

Gene Circuit Design and Construction

Having set the standards for biological components, the next step is to develop minimum functional components from the standard parts. They are typically gene circuits that express particular proteins. These single circuits are connected together to construct a gene regulatory network. Network motifs, such as cascades, feedforward, and feedback, are commonly used for the design of artificial regulatory network (and they are also

commonly found in natural biological systems) [7]. Cascades are series-connected gene circuits that output (protein) of an upstream gene circuit regulates expression of its immediately downstream gene. Feedforward motif involves a master regulatory gene that influences downstream genes through noncircular pathways. When the feedforward loop is introduced in a network, for example, it is possible to construct a network that shows a nonlinear non-monotonic change in the output. For example, a sigmoid function with a steep transition can be implemented with two gene circuits connected by negative (inhibitory) feedforward motif [8]. Basu and coworkers implemented a pulse-generator system using the motif [9]. When the input stimulus is monotonously changed from low to high, the output of the pulse-generator network changes from low to high and then back to low. In addition, the pulse amplitude and delay reflect not only the input concentration but also its rate of change. Feedback is a common biological regulatory motif not only for intracellular gene networks but also for intercellular control (e.g., negative feedback regulation of hormonal control). Feedback loop allows various complex regulations over expression of genes: noise reduction [10], expression level control [11], bistable gene toggle switch [12], autonomous genetic oscillator [13], etc. In principle, any artificial regulatory scheme can be implemented with the combination of the above network motifs.

Although standardization and other engineering concepts, such as abstraction, modularity, and reliability, greatly contribute to the speed and tractability of gene circuit design, one should note the liquid nature of biological systems and components. In contrast to solid electrical components, biological components, DNA, RNA, and proteins, all operate in the liquid solution which provide an appropriate “context” (e.g., pH, temperature, materials, energy, etc.) for the systems to function. This cellular context dependence and sensitivity make it difficult for biological components to be modular and interchangeable. A gene circuit from naturally occurring system is likely to be optimized to a specific cellular context through evolution. Thus, it may not function when it is

placed in an artificial context. Ron Weiss pointed out the need to understand “how the function of a module or an entire biological system can be derived from the function of its component parts “and establish” the biological rules of composition” in order to build biological components and modules [2].

Minimal Life

Craig Venter and his coworkers took the synthetic gene design to its extreme. They proposed to synthesize a complete genome from scratch and create “artificial life.” In 2003, they have synthesized the whole genome of Phi-X 174, a 5386 base pair bacteriophage, from synthetic oligonucleotides, which took them only 14 days to complete the whole process [14]. The synthesized genome was injected into *E. coli* bacteria and confirmed to be infectious. After the success in synthesized virus, they took a further step to create a synthesized organism that has a minimal genome. A bacterial genome of *Mycoplasma genitalium*, which is one of the smallest known genomes in any living organisms (517 genes, 580,000 DNA base pairs), was chosen as a base for the synthesis. They removed genes in the genome that are considered redundant and injected into a *M. capricolum* cell in which the original genome was removed in order to test if the cell with the modified genome can be alive. In 2010, they reported that the size of the genome was eventually reduced to approximately one mega base pair that are minimal and sufficient to be considered alive. They called the cell *Mycoplasma laboratorium* which is the first artificial life, of which parent is a digital computer [15].

Another important aspect of minimal life in synthetic biology is “cells as chassis.” The context dependence of biological components requires the constant maintenance of appropriate cellular context. Artificially synthesized biological components and modules are not exception and therefore require cellular context and its maintenance to perform tasks. Apart from the synthetic genome mentioned above, thus, all the other parts of protein expression process (transcription, RNA processing, translation, protein folding,

amino-acid modifications) can be subjects of synthetic biology research [16]. The construction of a minimal cell from scratch, as a basic unit for proving cellular context, is also an active area in this sub-research field. Lipid vesicles (liposomes) are commonly used as minimal bioreactors to simulate biological cells surrounded by lipid bilayer membrane. Functions of biological cells, such as membrane division, fusion, protein expression, and self-replication, have been attempted to replicate inside vesicles [17]. In recent years, microfluidic techniques, called digital microfluidics, have been applied to the production of lipid vesicles [18]. These techniques allow high-throughput production and screening of protein expression in the microfluidic devices and are expected to facilitate the more precise control of experimental conditions.

Applied Biological and Chemical Synthesis

Practical applications of synthetic biology are expected to be quite enormous because one of ultimate goals in synthetic biology is to create synthetic organisms that perform commercially useful tasks. For example, Craig Venter's group developed the synthetic genome described above so that it could be served as a platform for future synthetic organisms on which new functions (i.e., gene circuits) can be added. Since 2009, Venter's company, Synthetic Genomics Incorporated (SGI), has been working with Exxon Mobil Corporation to develop synthetic algae that produce next-generation biofuel, such as ethanol or hydrogen [19]. They also claim it would be possible to synthesize organisms which can fix carbon dioxide using photosynthesis to mitigate climate change [20].

Another example is medical and pharmaceutical applications of synthetic biology. Microbes, typically bacteria and yeast, are engineered and used as "chemical factory" to produce chemical products that are difficult to synthesize artificially. This process inevitably involves metabolic pathway engineering of living cells and therefore requires management of complex and emergent

metabolic processes [21]. Keasling and coworkers developed synthetic bacteria for the antimalarial drug production [1, 22]. The drug, artemisinin, extracted from *Artemisia annua L* (sweet wormwood) is known to be the most effective drug for the treatment of malaria, but expensive to produce because of short supply of the wood. The engineered bacteria incorporate several genes originally from bacteria *E. coli* and yeast *S. cerevisiae* to produce artemisinic acid, a biogenetic precursor of artemisinin. First, MevT operon encodes enzymes that transform acetyl-CoA into mevalonate. Enzymes encoded by MBIS operon then transform mevalonate into farnesyl pyrophosphate (FPP). Additionally, genes for amorphadiene synthase (transforming FPP into amorphadiene) and oxidase and redox partners (amorphadiene into artemisinic acid) were introduced. Although *E. coli* has a native pathway, 1-deoxy-D-xylulose 5-phosphate (DXP) pathway, to produce FPP, it has been found that the engineered mevalonate pathway produces more amorphadiene than the native pathway. Artemisinic acid purified from bacterial extracts is transformed to artemisinin using established chemistry.

Cross-References

- Computational Systems Bioinformatics for RNAi
- Liposomes
- Nanotechnology Applications in Polymerase Chain Reaction (PCR)
- Structure and Stability of Protein Materials

References

1. Benner, S.A., Sismour, A.M.: Synthetic biology. *Nat. Rev. Genet.* **6**(7), 533–543 (2005)
2. Andrianantoandro, E., Basu, S., Karig, D.K., Weiss, R.: Synthetic biology: new engineering rules for an emerging discipline. *Mol. Syst. Biol.* **2**(1), 2006.0028 (2006)
3. The BioBricks Foundation. <http://bbf.openwetware.org/>
4. Knight, T.: Idempotent vector design for standard assembly of biobricks standard biobrick sequence

- interface. MIT Synthetic Biology Working Group Report, MIT Artificial Intelligence Laboratory. 1–11 (2003)
5. Endy, D.: Foundations for engineering biology. *Nature* **438**(7067), 449–453 (2005)
 6. Nielsen, P.E., Egholm, M.: An introduction to peptide nucleic acid. *Curr. Issues Mol. Biol.* **1**(1–2), 89–104 (1999)
 7. McDaniel, R., Weiss, R.: Advances in synthetic biology: on the path from prototypes to applications. *Curr. Opin. Biotechnol.* **16**(4), 476–483 (2005)
 8. Hooshangi, S., Thiberge, S., Weiss, R.: Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. *Proc. Natl. Acad. Sci. U. S. A.* **102**(10), 3581–3586 (2005)
 9. Basu, S., Mehreja, R., Thiberge, S., Chen, M.-T., Weiss, R.: Spatiotemporal control of gene expression with pulse-generating networks. *Proc. Natl. Acad. Sci. U. S. A.* **101**(17), 6355–6360 (2004)
 10. Becskei, A., Serrano, L.: Engineering stability in gene networks by autoregulation. *Nature* **405**, 590–593 (2000)
 11. Nevozhay, D., Adams, R.M., Murphy, K.F., Josic, K., Balzsi, G.: Negative autoregulation linearizes the dose–response and suppresses the heterogeneity of gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **106**(13), 5123–5128 (2009)
 12. Gardner, T.S., Cantor, C.R., Collins, J.J.: Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403**(6767), 339–342 (2000)
 13. Judd, E.M., Laub, M.T., McAdams, H.H.: Toggles and oscillators: new genetic circuit designs. *Bioessays* **22**(6), 507–509 (2000)
 14. Smith, H.O., Hutchison, C.A., Pfannkoch, C., Venter, J.C.: Generating a synthetic genome by whole genome assembly: x174 bacteriophage from synthetic oligonucleotides. *Proc. Natl. Acad. Sci. U. S. A.* **100**(26), 15440–15445 (2003)
 15. Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N., Chuang, R.-Y., Algire, M.A., Benders, G.A., Montague, M.G., Ma, L., Moodie, M.M., Merryman, C., Vashee, S., Krishnakumar, R., Assad-Garcia, N., Andrews-Pfannkoch, C., Denisova, E.A., Young, L., Qi, Z.-Q., Segall-Shapiro, T.H., Calvey, C.H., Parmar, P.P., Hutchison, C.A., Smith, H.O., Venter, J.C.: Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329**(May), 52–56 (2010)
 16. Forster, A.C., Church, G.M.: Towards synthesis of a minimal cell. *Mol. Syst. Biol.* **2**(45), 45 (2006)
 17. Luisi, P.L., Ferri, F., Stano, P.: Approaches to semi-synthetic minimal cells: a review. *Naturwissenschaften* **93**(1), 1–13 (2006)
 18. Theberge, A.B., Courtois, F., Schaefer, Y., Fischlechner, M., Abell, C., Hollfelder, F., Huck, W. T.S.: Microdroplets in microfluidics: an evolving platform for discoveries in chemistry and biology. *Angew. Chem. Int. Ed.* **49**(34), 5846–5868 (2010)
 19. Synthetic Genomics Inc. Press Release <http://www.syntheticgenomics.com/media/press/71409.html>. Accessed 14 July 2009
 20. ETC group report, Extreme Genetic Engineering: An Introduction to Synthetic Biology, <http://www.etcgroup.org/content/extreme-genetic-engineering-introduction-synthetic-biology> (2007) Accessed 7 May 2015
 21. Khosla, C., Keasling, J.D.: Metabolic engineering for drug discovery and development. *Nat. Rev. Drug Discov.* **2**(12), 1019–1025 (2003)
 22. Keasling, J.: Synthetic biology for synthetic chemistry. *ACS Chem. Biol.* **3**(1), 64–76 (2008)

Synthetic Genomics

► Synthetic Biology

Synthetic Lubricants

► Boundary Lubrication

Systems Level Data Mining for RNAi

► Computational Systems Bioinformatics for RNAi

