Chapter 1

Some Trends in Chem(o)informatics

Wendy A. Warr

Abstract

This introductory chapter gives a brief overview of the history of cheminformatics, and then summarizes some recent trends in computing, cultures, open systems, chemical structure representation, docking, de novo design, fragment-based drug design, molecular similarity, quantitative structure–activity relationships (QSAR), metabolite prediction, the use of phamacophores in drug discovery, data reduction and visualization, and text mining. The aim is to set the scene for the more detailed exposition of these topics in the later chapters.

Key words: History of cheminformatics, Cheminformatics, Chemical structures, 2D searching, 3D searching, Similarity, Protein–ligand docking, Virtual high throughput screening, De novo design, Fragment-based drug design, Ligand-based drug design, QSAR, Computing infrastructure, Pharmaceutical industry, Open systems, Metabolite prediction, Pharmacophore, Visualization, Text mining

1. Introduction

Despite the fact that chem(o)informatics started to emerge as a distinct discipline in the late 1990s, its practitioners have still not alighted on a title for it [1, 2]: the terms "cheminformatics," "chem(o)informatics," "chemical informatics," and "chemical information" have all been used, but "cheminformatics" is the most commonly used name [3]. The *Journal of Chemical Information and Modeling* (formerly the *Journal of Chemical Information and Computer Sciences*), the core journal for the subject [2], still uses the term "chemical information" despite the fact that until recently chemical information was a discipline more associated with databases and "research information" [4, 5].

Not only is there no agreement on the name of the discipline, but also there is no agreed definition of what is involved [1, 2].

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_1,

Despite all this, practitioners do belong to a certain community, and one that is truly international in nature. It is possible to distinguish the learned journals that publish papers in the field [2, 6]. The *Journal of Chemical Information and Modeling* is the core journal but many significant papers are published in journals whose principal focus is molecular modeling or quantitative structure–activity relationships (QSAR) or more general aspects of chemistry [2]. Typical specialized journals are the *Journal of Medicinal Chemistry*, the *Journal of Computer-Aided Molecular Design*, the *Journal of Molecular Graphics and Modelling*, and *QSAR & Combinatorial Science*, but cheminformatics papers also appear in more broadly based journals such as *Drug Discovery Today* and *Current Opinion in Drug Discovery and Development*. The textbooks most commonly used in cheminformatics courses are those by Leach and Gillet [7] and Gasteiger and Engel [8]. Books edited by Bajorath [9] and by Oprea [10] are also recommended. Schneider and Baringhaus have produced a more recent text book [11].

The current book, *Chemoinformatics and Computational Chemical Biology*, is a successor to Bajorath's 2004 book *Chemoinformatics: concepts, methods, and tools for drug discovery* [9]. Its chapters cover such a wide range of topics that it is not possible in this introductory article to give a detailed analysis of trends in each of the fields. Instead, some general trends will be highlighted, and illustrated with a selected and by no means comprehensive set of examples. First, there is a very brief history of selected fields in cheminformatics up to about the year 2000, and after that newer developments in each field are considered in separate sections.

2. History

Several useful histories have been published recently [12–16]. Chemical structure handling is a mature technology [17–20]. Chemical databases, structure and substructure searching in 2D [17–21], reaction retrieval [17, 22], generation of 3D structures from 2D structures [23–26], 3D searching [27–34], similarity search of 2D or 3D structures [32, 35], and retrieval of generic ("Markush") structures [36, 37] are well documented.

Reaction systems can be subdivided into reaction retrieval systems (such as REACCS and its successors from Symyx Technologies, CASREACT from Chemical Abstracts Service, and Reaxys from Elsevier) and synthetic analysis programs [38]. Ott [22] further divides the latter class into synthesis design programs (such as LHASA, which relies on a knowledge base) and WODCA (now THERESA) [39], which performs retrosynthesis in a logic-oriented fashion); reaction prediction programs, such as Ugi's

IGOR [40, 41], Gasteiger's EROS [42], and Jorgensen's CAMEO [43]; and mechanism elucidation programs.

Programs for the generation of 3D structures from 2D structures [23–26, 44, 45] spurred on development of 3D structure methods because at that time only a limited number of experimentally determined 3D structures were available in databases. The history of crystallographic databases goes back to the early 1970s, but it has taken many years for them to grow. The Cambridge Structural Database [46–49] is the world repository of small molecule crystal structures. By January 2009 it contained 469,611 structures. The Protein Data Bank (PDB) began as a grassroots effort in 1971. It has grown from a small archive containing a dozen structures to a major international resource for structural biology containing more than 40,000 entries [50]. It contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies.

If the structure of a drug target is known (e.g. is in the PDB) it can be used in structure-based drug design. In computer-aided drug design four situations can arise. In the first case, if the receptor structure is unknown and there are no known ligands, computational techniques can be used to design collections of compounds with diverse structures for high throughput screening (HTS). In another case, the structure of the target or receptor is known, and the structure of the ligand is known. In this case protein-docking algorithms can be used to place candidate compounds within the active site of the target and rank-order them. In the third situation, the structure of the receptor is known, but the structure of the ligand is unknown; de novo design techniques can propose new ligands that are complementary to the active site.

In the final case, where the receptor structure is unknown, but there are ligands of known structure, ligand-based drug design is carried out. Similarity-based and machine learning-based "virtual screening" can be used, for example, or starting with a collection of molecules of known activity, the computational chemist can develop either a QSAR model or a 3D pharmacophore hypothesis that is converted into a search query. Scientists then use the search query to search a 3D database for structures that fit the hypothesis, or they use the QSAR model to predict activities of novel compounds.

Early programs for pharmacophore mapping (also called elucidation) [51] included DISCO [52], Catalyst [53, 54], and Genetic Algorithm Superimposition Program GASP [55–57]. Catalyst is still in regular use today. It has two components: HypoGen, a regression-like method for generating and optimizing hypotheses from 15–30 compounds with a range of potencies and HipHop for finding key features and producing alignments from a small set of potent compounds. Pharmacophore tools must allow for the fact that many compounds are flexible and can

assume multiple conformations. The exploration of multiple conformations (and often large numbers of them) can be tackled by generating and storing multiple representative conformations (as in Catalyst) or by exploring conformational space "on the fly."

GASP employs a genetic algorithm (GA) for the superimposition of sets of flexible molecules. Molecules are represented by a chromosome that encodes angles of rotation about flexible bonds and mappings between pharmacophore-like features in pairs of molecules. The molecule with the smallest number of features in the data set is used as a template, onto which the remaining molecules are fitted. The fitness function of the GA is a weighted combination of the number and the similarity of the features that have been overlaid in this way; the volume integral of the overlay; and the van der Waals energy of the molecular conformations defined by the torsion angles encoded in the chromosomes.

In ligand-based drug design, an alternative to the use of pharmacophores is QSAR. The QSAR method involves the conversion of molecular structures into mathematical descriptors that capture the properties of molecules that are relevant to the activity being modeled; selecting the best descriptors from a large set; mapping those descriptors onto the activities; and validating the model to determine how predictive it is and how well it can be extrapolated to molecules not in the training set used to generate the model [58–62]. Descriptors may be calculated from 2D or 3D structures. In the widely used Comparative Molecular Field Analysis (CoMFA) method [63], a 3D structure is surrounded by an array of grid points. At each point outside the molecule a probe atom or functional group is used to calculate steric, electrostatic, and sometimes, lipophilic fields at that point. One disadvantage is that an alignment rule is needed to superimpose molecules in the training set. A related method is comparative molecular similarity indices analysis (CoMSIA) [64]. FlexS [65] can be used to align molecules in 3D and prepare compounds for 3D QSAR analysis.

Deficiencies in absorption, distribution, excretion, metabolism (ADME) characteristics are the leading causes of attrition during drug development [66]. Prediction of toxicology is particularly difficult because of the variety of biological processes involved, but much research has been carried out in this field [67, 68]. Lipinski's "rule of five" [69] is a widely used rule of thumb to predict "druglikeness." Lipinski's rule says that, in general, an orally active drug has no more than one violation of the following criteria: not more than five hydrogen bond donors, not more than 10 hydrogen bond acceptors, a molecular weight under 500 Da, and an octanol–water partition coefficient (Clog*P*) of less than 5.

Widespread adoption of HTS and chemical synthesis technologies in the 1990s led to a data deluge. Combinatorial chemistry

allows very large numbers of chemical entities to be synthesized by condensing a small number of reagents together in all possible combinations. A "chemical library" is a set of mixtures or discrete compounds made by one combinatorial reaction. As an indication of the size of the chemical space covered by one library, Cramer [70] quotes the reaction of 4,145 commercially available diamines with R groups from 68,934 acylating reagents, cleanly displaceable halides, etc. (used twice), giving a library 2.0×10^{13} compounds. Compare this number with the 50 million known compounds in the CAS Registry. Cramer calculates that it would take 60,000 years of screening at the rate of 1 million per day, to test 2.0×10^{13} compounds. In addition, random screening has proved too expensive, its hit rate is low, false positives may be a problem, and expensive compounds are consumed. At the same time, developments in hardware and software during the 1990s meant that larger amounts of 3D structural information could be processed and allowed new methodological approaches to computer-aided drug discovery [32, 71]. All of this proved fruitful for progress in computational chemistry.

Selection of those compounds most likely to be hits was of considerable interest. It has been said (in a statement often attributed to David Weininger) that there are 10^{180} possible compounds, 10^{18} likely drugs, 10^{7} known compounds, 10^{6} commercially available compounds, 10^{6} compounds in corporate databases, 10^{4} compounds in drug databases, 10^{3} commercial drugs, and 10^{2} profitable drugs. Early library design efforts [14, 35, 72–74] involved selecting diverse subsets by clustering, dissimilarity-based selection, partitioning/cell-based approaches, or optimization-based methods [75]. Initially, diverse or focused libraries were designed based on descriptors for the reagents, since this required less computation, but later it was shown that product-based design produces more diverse libraries [76]. Moreover, diversity (or similarity, with design focused on certain specific chemical series) is not the only criterion for compound selection. Other factors such as druglikeness and synthetic accessibility need to be considered. This need led to progress in the field of multiobjective library design [77, 78].

To be considered for further development, lead structures should be members of an established SAR series; and should have simple chemical features, amenable to chemical optimization, a favorable patent situation, and good ADME properties. By analyzing two distinct categories of leads, those that lack any therapeutic use (i.e. "pure" leads), and those that are marketed drugs themselves but have been altered to yield novel drugs, Oprea and colleagues [79, 80] have shown that the process of optimizing a lead into a drug results in more complex structures. Hann and co-workers have studied molecular complexity [81]

and shown that less complex molecules are more common starting points for the discovery of drugs. These studies of leadlikeness and druglikeness have contributed to a trend in designing libraries of less complex, leadlike molecules, to leave scope for the almost inevitable increase in size and complexity during the optimization process.

Libraries may be designed for biological screening purposes, but computers may also be used to predict the results of screening in a process called "virtual screening." Nowadays, the term "virtual high throughput screening" is sometimes equated with protein–ligand docking, but other methods for virtual HTS have been developed, including identifying drug-like structures [82, 83], 2D similarity [35, 84], pharmacophore elucidation, and the use of 3D pharmacophores in 3D database searching [27–30, 32, 71]. An algorithm for docking small molecules to receptors (later to become the DOCK program) was published by Kuntz et al. [85] as long ago as 1982. Other programs started to appear in the late 1990s [55–57, 86]. Many factors, including an increase in the power of computers, have made docking increasingly popular of late.

3. Computers and Computing Environments

It is likely that a man was put on the moon in 1969 using a computer less powerful than today's typical cell phone. The cheminformatics systems of the 1980s and 1990s were run on mainframes or "minis" less powerful than today's PC. The VAX 11/750 used to run the U.K. Chemical Database Service [87] in 1984 had a clock speed of 6 MHz, 2 Mb memory, 134 Mb fixed disk, and two 67 Mb exchangeable disk drives; by judicious sharing of peripherals, the cost was kept down to £100,000 (1984 price). Readers need only to look at the specification and cost of their own hardware to see the advances of the last 20 years. Nowadays, parallel code and grid computing are commonplace, and cyberinsfrastructure (e-science) has been established [88] as an enabling platform.

For applications needing more speed than the average CPU can provide, field programmable gateways (FPGAs), graphics processing units (GPUs), and even gaming devices such as Microsoft's Xbox gave been explored. For example, the so-called Lightning version of SimBioSys' eHiTS ligand docking software [89] has been run on the Sony PlayStation 3 (PS3) game console [90], or more specifically on a microprocessor architecture called the Cell Broadband Engine (Cell/B.E.) which powers the PS3. The Cell/B.E. enables the PS3 to speed up physics simulations so that

they can catch up with the 3D graphics rendering speeds of the system's GPUs. The IBM BladeCenter QS21 blade server is based on the same Cell/B.E. processor.

The emergence of the World Wide Web in 1992 brought about an information revolution. Now "Web 2.0" technologies [91–93] and the Semantic Web [94–97] are having an impact. "Web 2.0" is a badly defined term covering multiple collaborative technologies such as instant messaging, text chat, Internet forums, weblogs ("blogs"), wikis, Web feeds, and podcasts; social network services including guides, bookmarking, and citations; and virtual worlds such as Second Life [91]. Some of these technologies may seem to have little relevance to cheminformatics, but there are in fact some applications of interest, e.g. Pfizerpedia [98], the Pfizer wiki, use of which has reportedly increased exponentially.

In 2005, CAS introduced CAS Mobile for real-time interaction with CAS databases using wireless handheld devices. In 2009, the application ChemMobi was posted to the Apple App Store and can be downloaded, for free, to enable an iPhone to search both Symyx's Discovery Gate [99] and ChemSpider [100]. ChemMobi uses DiscoveryGate Web Service and the ChemSpider Web Service: Web Services are another feature of today's computing architectures. "Cloud computing" (information infrastructure, software, and services hosted on the Internet rather than on one's own computer) is in its infancy but is attracting much interest in the technology press.

Pipelining and workflow methods have long been used in bioinformatics but later started to impact cheminformatics [101]. The workflow paradigm is a generic mechanism to integrate different data resources, software applications and algorithms, Web services, and shared expertise. (There are minor differences between pipelining and workflow.) Such technologies allow a form of integration and data analysis that is not limited by the restrictive tables of a conventional database system. They enable scientists to construct their own research data processing networks (sometimes called "protocols") for scientific analytics and decision making by connecting various information resources and software applications together in an intuitive manner, without any programming. Therefore, software from a variety of vendors can be assembled into something that is the ideal workflow for the end user. These are, purportedly, easy-to-use systems for controlling the flow and analysis of data. In practice, certainly in chemical analyses, they are not generally used by novices: best use of the system can be made by allowing a computational chemist to set up the steps in a protocol then "publish" it for use by other scientists. KNIME and Pipeline Pilot are the systems most familiar to computational chemists but Kepler, Taverna, SOMA, and InforSense are also worthy of note [101].

4. Influence of Industry

Much of the research in this subject area is carried out in industry and the majority of the applications are written in industry [6], and in particular the pharmaceutical industry [102]. Thus, it is not surprising that a significant number of the authors in this book have, or have had, industrial affiliations. The pharmaceutical industry, however, currently faces unprecedented problems, including the increasing cost of R&D, the decreasing number of new chemical entities, patent expiry on "blockbuster" drugs between 2008 and 2013, and pressure to reduce drug prices. This has led to new strategies and increased interest in translational research, pharmacogenomics, biomarkers, and personalized medicine. Cheminformatics is likely to change in response to these changes.

Industry and academia have always collaborated in cheminformatics projects but a new trend is the availability in academia of biological assay data: the sort of data that used to be largely held in proprietary systems in the pharmaceutical industry. Data from the Molecular Libraries Screening Centers in the United States are available through PubChem [103, 104]. The National Institutes of Health (NIH) are also launching the Therapeutics for Rare and Neglected Diseases (TRND) program [105] which creates a drug development pipeline within the NIH and is specifically intended to stimulate research collaborations with academic scientists working on rare illnesses. The cultures of bioinformatics and cheminformatics have tended to differ when it comes to open applications and sharing of data: cheminformaticians are the more likely to use proprietary software and data.

5. Open Systems

The pharmaceutical industry in general has not been keen to embrace open source software, although some individuals are enthusiastic [106, 107]. Limited attempts have been made at collaboration (the commercially available organic chemicals alliance in the 1980s was a success in some respects, and LHASA Limited continues to this day), but the current trend to open systems has led to renewed interest in collaborating in areas which do not give competitive advantage [108, 109].

The Pistoia Alliance [108] started with an initial meeting (in Pistoia, Italy) where proponents at GlaxoSmithKline, AstraZeneca, Pfizer, and Novartis outlined similar challenges and frustrations in the IT and informatics sector of discovery. The advent

of Web Services and Web 2.0 allows for decoupling of proprietary data from technology. A service orientated approach allows for these types of pre-competitive discussions. The primary purpose of the Pistoia Alliance is to streamline non-competitive elements of the pharmaceutical drug discovery workflow by the specification of common business terms, relationships, and processes. There is a vast amount of duplication, conversion, and testing that could be reduced if a common foundation of data standards, ontologies, and Web Services could be promoted and ideally agreed within a non-proprietary and non-competitive framework. This would allow interoperability between a traditionally diverse set of technologies to benefit the healthcare sector. The Pistoia Alliance was officially launched in February 2009. Additional members include ChemITment, ChemAxon, Accelrys, Edge Consultancy, BioXPR, GGA, Lundbeck, Bristol Myers Squibb, Roche, KNIME, Rescentris, and DeltaSoft.

A case study is the LHASA Web Service. Before Pistoia involvement, LHASA's DEREK [110] software did not have a Web Service; it had only a Windows-based API. Each company had created or adapted its own LHASA interface, which was subject to change as LHASA updated the product. Each company had done much the same interfacing. There was no consistent approach to a Web Service. After Pistoia involvement, there is a single LHASA DEREK Web Service available to all customers.

Similar trends are occurring in bioinformatics [109]. In cheminformatics, there is now considerable interest in "open source, open standards, open data," and "reusable chemistry" [91, 92, 94, 95]. Large, chemical structure databases with physical property and activity data such as ZINC [111], PubChem [103], ChemSpider [100], eMolecules [112], and the NIH/CADD Chemical Structure Lookup Service [113] have become freely available [114]. These databases contain millions of molecules but even larger databases of virtual molecules have been produced [115, 116]. The Collaborative Drug Discovery [117] portal enables scientists to archive, mine, and collaborate around preclinical chemical and biological drug discovery data through a Web-based interface. Free exchange of chemical structures on the Web has become easier with the emergence of open standards for identification.

6. Chemical Structure Representation

The Morgan algorithm [118] underpins many of the systems in use today, and is the basis of the CAS REGISTRY database. It identifies atoms based on an extended connectivity value; the atom with the highest value becomes the first atom in the

name, and its neighbors are then listed in descending order. Ties are resolved based on additional parameters, for example, bond order and atomic number. The original Morgan algorithm did not handle stereochemistry; the Stereochemically Extended Morgan Algorithm (SEMA) was developed to handle stereoisomers [119]. The Newly Enhanced Morgan Algorithm (NEMA) [120] produces a unique name and key for a wider range of structures than SEMA. It extends perception to non-tetrahedral stereogenic centers, it supports both 2D and 3D stereochemistry perception, and it does not have an atom limit. The CAS Registry system, SEMA and NEMA are proprietary.

Systems such as ChemSpider use the International Union of Pure and Applied Chemistry (IUPAC) International Identifier (InChI) to register chemical structures. IUPAC developed InChI as a freely available, non-proprietary identifier for chemical substances that can be used in printed and electronic data sources, thus enabling easier linking of diverse data compilations and unambiguous identification of chemical substances. IUPAC decided to tackle this problem because the increasing complexity of molecular structures was making conventional naming procedures inconvenient, and because there was no suitable, openly available electronic format for exchanging chemical structure information over the Internet. The goal of the IUPAC International Chemical Identifier (InChI) is to provide a unique string representing a chemical substance of known structure, independent of specific depiction, derived from conventional connection table, freely available, and extensible [121–123].

The InChI project was initially undertaken by IUPAC with the cooperation of National Institute for Standards and Technology (NIST). Steps 1–3 of the canonical numbering for InChI are done using an algorithm modified from that of McKay [124]. In 2009, a standard version of InChI and the InChIKey were released. InChIKey is a fixed length condensed digital representation of the identifier which facilitates Web searching, previously complicated by unpredictable breaking of InChI character strings by search engines. It also allows development of a Web-based InChI lookup service; permits an InChI representation to be stored in fixed length fields; and makes chemical structure database indexing easier. InChI has been used in chemical enhancement of the Semantic Web [94].

Like InChI, the SMILES language [125, 126] allows a canonical serialization of molecular structure. However, SMILES is proprietary and unlike InChI, it is not an open project. This has led to the use of different generation algorithms, and thus, different SMILES versions of the same compound have been found. InChI is not a registry system such as that of CAS; it does not depend on the existence of a database of unique substance records

to establish the next available sequence number for any new chemical substance being assigned an InChI.

Having looked at recent trends in the basics of cheminformatics (hardware, infrastructures, and cultural issues) we will now consider drug design technologies one by one, starting with protein–ligand docking, an approach which has proved increasingly popular in recent years.

7. Docking

Receptor–ligand docking [127] is a computational procedure that predicts the binding mode and affinity of a ligand to a target receptor. In this method, each ligand in a database is docked into the active site of the receptor. Docking programs require an algorithm that can explore a very large number of potential docking conformations and orientations for each ligand. These programs generate a series of 3D models that predict the way in which each small molecule will bind to the targeted receptor. Docking programs also include a scoring function that quantitatively ranks the ligands according to their binding affinity for the receptor. Although scoring functions are meant to determine which compounds are more likely have a higher affinity for the target molecule, in practice, these functions do not always accurately rank ligands.

More than 60 docking programs (e.g., Autodock, DOCK, eHiTS, FlexX, FLOG, FRED, Glide, GOLD, Hammerhead, ICM, Surflex-Dock) and more than 30 scoring functions [128] have been recorded. Many different docking methods are available including fast shape matching, distance geometry, genetic algorithms, simulated annealing, incremental construction, and tabu search. Scoring has been carried out with force fields, empirical schemes, potentials of mean force, linear interaction energies, or other molecular dynamics.

In theory, docking methods provide the most detailed description of all virtual screening techniques. Ideally, ligand–receptor docking methods describe how a compound will interact with a target receptor, what contacts it will make within the active site, and what binding affinity it will have for the receptor. Although pharmacophore-based searching can select compounds possessing approximately the desired 3D characteristics, this approach does not describe how well matched the molecule and the receptor are in terms of shape or other properties and does not predict binding affinity or rank the selection in any way. Also, while statistical selection methods are excellent at finding compounds with structural similarities to known ligands, docking methods are better suited to finding novel structural types.

Docking, however, is such a computationally demanding exercise that, in practice, its limitations or approximations are inevitable. For example, although many current docking programs account for some degree of ligand flexibility, sampling [128] of the many possible conformations of each ligand in a database must be limited if the search is to be completed in a reasonable length of time. Also, when a ligand binds to its receptor, the receptor may also undergo conformational change. Docking programs that use rigid receptor models do not account for these changes.

Another important issue for researchers using docking as a virtual screening strategy is the accuracy of the predicted binding modes and affinities. Studies have shown that while some docking programs can reproduce crystal structures well, they generate scores that do not correlate with measured IC_{50} values. Of particular concern is the ability of current scoring functions to rank ligands correctly. To dock all the molecules contained in very large database, a scoring function must be simple, very fast, trained across a wide variety of proteins, and derived from a physically reasonable equation. Researchers have devised various scoring functions; however, no one has yet derived a truly accurate, broadly applicable method. Scoring functions used in virtual screening often work well for some targets but not as well for others. To overcome this hurdle, some programs allow more than one function to be employed with a single program. Consensus scoring, which combines several scoring methods, has been found to be superior to the use of a single function in some cases [129].

Another strategy is to use a simple function to discriminate between alternative binding geometries for a single ligand and combine it with more elaborate calculations to predict binding affinities. "Physics-based" methods to estimate binding affinity are computationally expensive but more accurate [102, 130, 131], but here we are moving into the realm of theoretical chemistry as opposed to cheminformatics.

The many docking programs currently available are usually judged (apart from speed) in terms of pose accuracy and enrichment (the ratio of the observed fraction of active compounds in the top few percent of a virtual screen to that expected by random selection). A large number of comparative and validation studies have been carried out [128, 132, 133]; progress has been aided by the availability of useful data sets such as ZINC, which contains over 13 million purchasable compounds in ready-to-dock, 3D formats [111], and the Directory of Useful Decoys (DUD), derived from ZINC, for benchmarking virtual screening [134, 135]. DUD contains a total of 2,950 active compounds against a total of 40 targets, and for each active, 36 "decoys" with similar physical properties (e.g., molecular weight and calculated Log*P*) but dissimilar topology. Data from the World of Molecular

Bioactivity (WOMBAT) database have also been used in conjunction with DUD [136].

Carrying out a valid comparative study is fraught with difficulties [128, 132, 137, 138]. Factors include the versions of the programs, the settings employed, fine tuning of parameters, quality of the data sets, preparation of receptors and ligands, and the criterion used for measuring accuracy. Using the root mean-square deviation (RMSD) as a criterion has been questioned [139]. There is also debate about enrichment factors and ROC curves [140, 141]. Boltzmann-enhanced discrimination of receiver operating characteristic (BEDROC) [142] and robust initial enhancement (RIE) [143] have been proposed as alternatives to tackle the "early recognition" problem.

Several authors have found docking methods to be no more effective, or indeed less effective, than ligand-based virtual screening methods based on 3Dshape matching or 2D descriptors [144–146]. Several direct comparisons of docking with the shape-based tool ROCS [147] have been conducted using data sets from some recent docking publications. The results show that a shape-based, ligand-centric approach is more consistent than, and often superior to, the protein-centric approach taken by docking [144].

One way to improve a program's performance is to provide additional information such as pharmacophore(s) to orient the docking [148]. Self-docking is a good indication of a program's ability to identify native poses amongst several others, but provides little information about the accuracy in a real drug discovery scenario. For medicinal chemists, the best indicator of a program's accuracy is its ability to identify novel compounds that are then experimentally confirmed [131]. Apart from the need for better scoring functions, challenges remaining in protein–ligand docking include balancing speed and accuracy in conformational sampling; protein flexibility; loss of entropy; and protein desolvation [128, 149].

8. De Novo Design

De novo design involves the design of inhibitors from scratch given the target binding site. Typical programs include LUDI [150], SPROUT [151, 152], and BOMB [102, 131, 153]. More than 20 programs were reported in the literature in the early 1990s, but there was limited uptake of such systems because they designed molecules which were not easy to synthesize. De novo design is now popular again, but there is more consideration of generating molecules that are synthetically accessible [154] and which also represent non-obvious structural transformations.

Scores based on molecular complexity and retrosynthetic analysis are used to prioritize structures during generation. In these systems, molecular transformations are driven by known reactions. Synthetic feasibility is implicit in the rules, often based on a limited number of hand-picked reactions (typically derived from retrosynthetic analysis), and atom mapping is required. Gasteiger's team has devised a scoring method that rapidly evaluates synthetic accessibility of structures based on structural complexity, similarity to available starting materials, and assessment of strategic bonds where a structure can be decomposed to obtain simpler fragments. These individual components are combined to give an overall score of synthetic accessibility by an additive scheme [152, 155, 156]. Gillet and co-workers have developed a knowledge-based approach to de novo design which is based on reaction vectors that describe the structural changes that take place at the reaction center, along with the environment in which the reaction occurs [157].

9. Fragment-Based Drug Design

In fragment-based drug design (FBDD), molecules having a molecular weight of approximately 110–300 Da (smaller than the usual drug molecule) are used in structure-based drug design. Interest in this approach began in the mid-1990s when workers at Abbott, using "SAR by NMR," proved that meaningful SAR and stable binding modes could be observed even with mM ligands [158]. At around that same time, X-ray crystallography was beginning to be used to map out "hot spots" in protein binding sites [159]. Thus, FBDD was born [160–162]. Both NMR and X-ray analyses provide structural information about the binding site of a hit. A concurrent theme is the pursuit of leadlikeness rather than druglikeness [79–81] discussed in Subsection 2.

FBDD has two main advantages. Firstly, chemical space can be more efficiently probed by screening collections of small fragments rather than libraries of larger molecules: HTS of a million compounds covers only a tiny proportion of the available chemical space of 10^{60} or more compounds, whereas virtual screening of 10,000 fragments covers a much higher *proportion* of chemical diversity space. Less complex molecules should also show higher hit rates [81, 163]. The second major concept of FBDD concerns binding efficiency. The ligand efficiency (LE) measure, binding energy divided by the number of heavy atoms [164], and other measures [165–167] have been devised. The "Astex Rule of 3" (three or fewer hydrogen bond acceptors, three or fewer hydrogen bond donors, and CLog$P \leq 3$) has been suggested for selecting suitable fragments [168]. Most corporate collections contain

molecules that have been optimized for historical targets. HTS sometimes fails to find hits that interfere with new targets, such as protein–protein interaction surfaces. FBDD is robust for novel and difficult target classes although the success rate is likely to be dramatically lower than for easier targets. FBDD can also find new binding modes and allosteric sites.

Once hits have been found they must be fully characterized before crystallography or SAR studies are carried out; there may be many false positives. Surface plasmon resonance (SPR) and NMR can be used to eliminate non-specific binders. It is usually assumed that determining the X-ray structure of the fragment and target is an essential next step, but NMR and modeling may be used if an X-ray structure cannot be found. Structure-based drug design can then be used in optimization of the fragment [169].

NMR is still the commonest way of finding leads. Unfortunately, both NMR and X-ray crystallography screening require high concentrations of both fragments and target: the fragments must be soluble at high concentrations. SPR has the advantage of providing quantitative dynamics data on the binding interaction, such as binding constants, which are complementary to the structural information from X-ray and NMR screens. Some companies have also used high concentration bioassays, thermal methods, mass spectrometry (MS), and MS plus tethering with extenders (small molecules that bind to an active-site cysteine and contain a free thiol) [170]. Orthogonal validation (using two or even three assay methods in parallel) is a fairly new trend.

It has been claimed that a fragment screen provides a rapid and reliable means of interrogating a protein target for drugability before investing in further discovery research [169]. Drugability score (DScore) calculated by Schrödinger's SiteMap has allowed targets of low and high hit rates to be differentiated [163].

There have been a number of reports of the use of high-concentration screening or "reduced complexity" screening on compound collections that are a hybrid of a true fragment library and of a typical HTS collection. The upper molecular weight limit, for example, may be up to 350 Da. The available subset of molecules that can be screened within a corporate collection is thus increased, and these larger molecules, will, if active, be detectable in a HTS campaign simply by screening at a higher than normal concentration. On the other hand it has been argued that with this technique, a much larger library of leadlike compounds will be required to achieve a hit rate comparable to that observed for small fragments screened using very sensitive techniques at a higher concentration, and that at lower concentrations the smallest fragments will only be detectable if they have potency similar to that of the larger, more complex compounds being screened [167]. Publications describing the design and characterization of fragment libraries are becoming more common [167, 171].

One problem is that fragment-based approaches identify and characterize only "hot spots," i.e. the regions of a protein surface that are major contributors to the ligand binding free energy. Unfortunately, many binding sites in the active site that are responsible for target specificity and/or selectivity are not included in these "hot spots." Fragment screening finds the most efficient binder in the smallest core but optimization is still needed.

Once a hit has been found, it is optimized into a lead by one of three approaches: linking, growing, or merging. Linking may appear to be a good approach (energetically) but it can be hard to find second site fragments and affinity can be lost in conformational strain of the linker. It is usually more successful to grow fragments by structure-guided medicinal chemistry or grow by using the fragment binding motif to search for similar compounds that can be purchased (a method often called "SAR by catalog"). The merging approach has not been widely adopted. FBDD does not make drug design easier but it does offer more options. Optimization is still difficult and in some cases it is an intractable problem. Research continues on in situ fragment assembly techniques such as click chemistry, dynamic combinatorial library design, and tethering with extenders. A number of "success stories" have been published [169, 172], but as yet there is no "FBDD drug" on the market.

10. Molecular Similarity Analysis and Maximal Common Substructure

A very recent review covers a range of similarity methods in cheminformatics [173]; novel approaches to molecular similarity analysis have also been reviewed recently [174]. Some of the novel methods relate to QSAR and are discussed in the next section. Willett has reviewed advances in similarity based screening using 2D fingerprints [84]. Many different similarity coefficients have been used, including Tanimoto, cosine, Hamming, Russell-Rao, and Forbes. Willett's team has studied the use of data fusion methods: the similarity fusion approach is akin to consensus scoring in docking. In group fusion, not just one reference structure, but several structurally diverse reference structures are used [84]. The team has also worked on so-called turbo similarity searching which seeks to increase the power of the search engine by using a single reference structure's nearest neighbors.

Arguably, the most obvious way to compute the similarity between two 2D structures is to compare their maximal common substructures (MCS). An MCS is a single contiguously connected common subgraph of a molecular structure present in a specific

fraction of all molecules. Unfortunately, MCS isomorphism algorithms are extremely time-consuming [173]. Algorithms for finding an MCS include clique detection and clustering [175]. MCS is a core component of commercially available packages [176] such as Accelrys' Pipeline Pilot [177], the Simulations Plus ClassPharmer program [178], and ChemAxon's Library MCS clustering [179].

11. Quantitative Structure–Activity Relationships

Classical 3D QSAR methods, such as CoMFA, often provide accurate prediction of biological activity. Moreover, CoMFA models are interpretable, suggesting chemical changes that will improve biological activity. However, these methods can require time-consuming, expert preparation (molecular alignment, and conformer selection). Topomer CoMFA [180–182] minimizes the preparation needed for 3D QSAR analysis through an objective and consistent set of alignment rules. Topomer CoMFA can be used in conjunction with Topomer Search to identify the substituents and R-groups that are predicted to optimize the activity of compounds.

QSAR is no longer modeled in just one, two, or three dimensions [183]. 4D QSAR was an early approach to solving the alignment problem in 3D QSAR [184]. In 1D-QSAR, affinity is correlated with physicochemical properties such as pK_a and $\log P$; in 2D-QSAR it is correlated with 2D chemical connectivity; in 3D-QSAR with the three-dimensional structure; in 4D-QSAR with multiple representations of ligand conformation and orientation (as well as 3D structure); in 5D-QSAR with multiple representations of induced-fit scenarios (as well as with 4D concepts); and in 6D-QSAR with multiple representations of solvation models (as well as with 5D terms).

It might be assumed that the main objective of a QSAR study is to predict, for example, whether an untested compound will be active or inactive but, in practice, much work has been devoted to "explanatory" QSAR, relating changes in molecular structure to changes in activity, and only recently has there been considerable interest in predictivity. There are many reasons why models fail [185, 186], not least bad data, bad methodology, inappropriate descriptors, and domain inapplicability [187]. Significant issues concerning accuracy of prediction are extrapolation (whether the model can be applied to molecules unlike those in the training set) and overfitting [188]. Running cross-validation studies on the data is a reasonable check for overfitting but it is inadequate as a measure of extrapolation [189].

The outcome of a leave-one-out (LOO), or leave-many-out, cross-validation procedure is cross-validated R^2 (LOO q^2). The inadequacy of q^2 as a measure of predictivity was realized more than 10 years ago, in what has been referred to as "the Kubinyi paradox": models that give the best retrospective fit give the worst prospective results [190]. The "best fit" models are not the best ones in external prediction because internal predictivity tries to fit compounds in the training set as well as possible and does not take new compounds into account [191]. Even in the absence of real outliers, external prediction will be worse than fit because the model tries to "fit the errors" and attempts to explain them [186].

While a high value of q^2 is a necessary condition for high predictive power, it is not a sufficient condition. Tropsha and coworkers have argued that a reliable model should be characterized by both high q^2 and a high correlation coefficient (R^2) between the predicted and observed activities of compounds from a test set [192, 193]. They have proposed several approaches to the division of experimental data sets into training and test sets and have formulated a set of general criteria for the evaluation of the predictive power of QSAR models. Other reasons for overestimating q^2 are redundancy in the training set, or, in the case of nonlinear methods, the existence of multiple minima [193].

Doweyko [194] concludes that predictions can be enhanced when the test set is bounded by the descriptor space represented in the training set. Gramatica has discussed principles to define the validity and applicability domain of QSAR models [195], and in particular, emphasizes the need for external validation using at least 20% of the data. Validation is essential for application and interpretation of QSAR models [187, 196] and this necessity has been accepted by leading journals [197–199].

Researchers at Merck [189] have proposed a way to estimate the reliability of the prediction for an arbitrary chemical structure, using a given QSAR model, given the training set from which the model was derived. They found two useful measures: the similarity of the molecule to be predicted to the nearest molecule in the training set and the number of neighbors in the training set, where neighbors are those more similar than a user-chosen cut-off. Nevertheless, incorrect predictions of activity still arise among *similar* molecules even in cases where overall predictivity is high, because in Maggiora's well known metaphor, activity landscapes are not always like gently rolling hills, but may be more like the rugged landscape of the Bryce Canyon [200]. Even very local, linear models cannot account satisfactorily for landscapes with lots of "cliffs," and perfectly valid data points located in cliff regions may *appear* to be outliers, even though they are perfectly valid data points.

Following Maggiora's observations, there has been research into "activity landscape" characterization. The success of ligand-based virtual screening is much influenced by the nature of target-specific

structure–activity relationships, making it hard to apply computational methods consistently to recognize diverse structures with similar activity [174]. The performance of similarity-based methods depends strongly on the compound class that is studied, and approaches of different design and complexity often produce, overall, equally good (or bad) results. Moreover, there is often little overlap in the similarity relationships detected by different approaches, so alternative similarity methods need to be developed. SARs for diverse sets of active compounds or analog series are determined by the underlying "activity landscapes" [201].

On the basis of systematic correlation of 2D structural similarity and compound potency, Peltason and Bajorath have developed "a structure activity index (SARI)" that quantitatively describes the nature of SARs and establishes different SAR categories: continuous, discontinuous, heterogeneous-relaxed, and heterogeneous-constrained. Given a set of active compounds and their potency values, SAR Index calculations can estimate the likelihood of identifying structurally distinct molecules having similar activity [202].

Guha and van Drie have also studied activity cliffs [203, 204]. By use of a quantitative index, the structure–activity landscape index (SALI), they have identified pairs of molecules which are most similar but have the largest change in potency and hence form activity cliffs. They have shown how this provides a graphical representation of the entire SAR (where each node is a molecule, and each edge represents an activity cliff of varying magnitude), allowing the salient features of the SAR to be quickly grasped. Consensus activity cliffs [205] have also been described.

It has been said that a general feeling of disillusionment with QSAR has settled across the modeling community [206] but actually there is renewed interest in QSAR in the field of absorption, distribution, excretion, metabolism, and toxicity (ADMET); many regulatory laws including the new Registration, Evaluation, Authorization of Chemicals (REACH) legislation in Europe have prompted significant new activity [207–209]. Under REACH regulation, information on intrinsic properties of substances may be generated by means other than tests, provided that certain conditions are met, so animal testing can be reduced or avoided by replacing traditional test data with predictions or equivalent data. Integrated testing strategies, including in vitro assays, QSARs, and "read-across," can be used in a combined "non-testing" strategy, i.e. as an alternative to the use of animals. In read across, known information on the property of a substance is used to make a prediction of the same property for another substance that is considered similar. This avoids the need to test every substance for every endpoint, but there are conditions. QSARs are allowed under REACH if the method is scientifically

valid, the domain is applicable, the endpoint is relevant, and adequate documentation is provided [207, 208].

Consensus QSAR has been applied to models developed from different techniques and different data sets; the method often performs better than a single QSAR [210]. One study by Cronin's team, however, shows that the use of consensus models does not seem warranted given the minimal improvement in model statistics for the data sets in question [211]. The Food and Drug Administration (FDA) has used multiple commercially available programs, with the same data sets, to predict carcinogenicity [212]. The FDA has several reasons for using more than one QSAR software program. None of the programs has all the necessary functionalities, and none has 100% coverage, sensitivity, and specificity. All of the programs are complementary. The individual models made complementary predictions of carcinogenesis and had equivalent predictive performance. Consensus predictions for two programs achieved better performance, better confidence predictions, and better sensitivity. Consensus models from three different expert systems have also been used with some success in prediction of mutagenicity using the commercial system KnowItAll [213]. Consensus models have been tailored to a risk assessment scenario in AstraZeneca [214]. There are, however, disadvantages. Consensus models hide outliers, incorrect data, and interesting parts of the data set. They lack portability, transparency, and mechanistic interpretation.

12. Metabolite Prediction

Various rule-based and statistical methods have been used to predict metabolic fate. Gasteiger's team [215] has used descriptors of drugs metabolized by human cytochrome P450 (CYP) isoforms 3A4, 2D6, and 2C9 in model building methods such as multinomal logistic regression, decision tree, or support vector machine (SVM). This team also supplies the biochemical pathways database, BioPath [216]. Schwaighofer and co-workers [217] have developed machine learning tools to predict the metabolic stability of compounds from drug discovery projects at Bayer Schering. They concluded that Gaussian Process classification has specific benefits.

Another team [214] has used data mining methods to exploit biotransformation data that have been recorded in the Symyx Metabolite [218] database. Reacting center fingerprints were derived from a comparison of substrates and their corresponding products listed in the database. The metabolic reaction data were then mined by submitting a new molecule and searching for

fingerprint matches. An "occurrence ratio" was derived from the fingerprint matches between the submitted compound and the reacting center and substrate fingerprint databases. The method enables the results of the search to be rank-ordered as a measure of the relative frequency of a reaction occurring at a specific site within the submitted molecule.

The rule-based method, Systematic Generation of Metabolites (SyGMa) [219] predicts potential metabolites based on reaction rules derived from metabolic reactions that occur in man, reported in Symyx' Metabolite database [218]. An empirical probability score is assigned to each rule representing the fraction of correctly predicted metabolites in the training database. This score is used to refine the rules and to rank predicted metabolites. Another team [220] has used absolute and relative reasoning to prioritize biotransformations, since they argue that a system which predicts the metabolic fate of a chemical should predict the more likely metabolites rather than every possibility.

13. Pharmacophores

Three-dimensional pharmacophore methods in drug discovery have been reviewed very recently [221]. The technologies used in 3D pharmacophore modeling packages such as Accelrys' Catalyst, Chemical Computing Group's Molecular Operating Environment (MOE), Schrödinger's Phase, and Inte:ligand's LigandScout have also been reviewed recently [222]. Another method, PharmID, uses fingerprints of 3D features and a modification of Gibbs sampling to align a set of known flexible ligands, where all compounds are active [223]. A clique detection method is used to map the features back onto the binding conformations. The algorithm is able to handle multiple binding mode problems, which means it can superimpose molecules within the same data set according to two different sets of binding features.

Alignment of multiple ligands is particularly difficult when the spatial overlap between structures is incomplete, in which case no good template molecule is likely to exist. Pairwise rigid ligand alignment based on linear assignment (the LAMDA algorithm) has the potential to address this problem [224]. A version of LAMDA is embodied in the program named Genetic Algorithm with Linear Assignment for Hypermolecule Alignment of Datasets (GALAHAD) developed by Tripos in collaboration with Biovitrum and Sheffield University [225]. GALAHAD creates pharmacophore and alignment models from diverse sets of flexible, active compounds, generating models that are sterically, pharmacophorically, and energetically optimal. It supports partial

matching of features and partial coverage. By decoupling conformational searching from alignment the program frees scientists from the need to fit all ligands to any single template molecule. GALAHAD uses a multi-objective optimization (Pareto ranking) scoring function. A single Pareto run produces better, more diverse models than an entire series of GA runs using a range of fitness term weights [78]. Tripos also offers the Surflex-Sim [226] method of molecular alignment and virtual screening.

Other new programs have become commercially available. Chemical Computing Group's MOE contains a pharmacophore elucidator that takes a collection of actives and inactive analogs, generates all pharmacophores (there may be multiple correct answers to a pharmacophore elucidation procedure) and validates each to see which one is best. Schrödinger's Phase is a package of pharmacophore modeling tools that offers scientists control at each step (including pharmacophore scoring, QSAR building, and database screening) and enables users to modify existing feature definitions and create new features. Some programs may be able to uncover multiple binding modes. The ability to use larger data sets, the coverage of conformational space, and the ability to handle more flexible molecules are reported to be other advantageous features.

High quality data sets have driven progress in the field of protein ligand docking but there are fewer data sets for pharmacophore mapping. It has been suggested that too many algorithms could have been developed using the same data, hampering progress in the field [227]. The Patel data set [228] has been used to evaluate Catalyst [53, 54], DISCO [52] and GASP [55–57] and has been refined [225] in the development of GALAHAD and in the Multi-Objective Genetic Algorithm (MOGA) program [77, 78]. MOGA looks for solutions which are compromises between three objective functions: energy, volume overlap, and feature score. It is now being tested on a new data set, the "Taylor" data set, which aims to cover 25–35 proteins, and is carefully compiled by an expert in the field. Protonation states are checked, and ligand geometries and electron density in the binding site are being checked. A detailed analysis of crystal structures is carried out to elucidate pharmacophore points. Each of the 10 protein targets has a minimum of two ligands and a maximum of 16 ligands [227].

Clark has discussed synergy between ligand-based and structure-based methods [229]. Unfortunately, ligand binding often induces structural changes that significantly reduce the usefulness of apoprotein structures for docking and scoring. In such cases it is often better to dock into the binding site of a ligand–protein complex from which the ligand has been extracted *in silico*. Even when a native protein structure is suitable for docking, ligands can

provide critical information about the location of the relevant binding site. Moreover, interactions with specific binding site residues illuminated by bound ligands have been successfully used to direct docking and to tailor scoring functions to specific target proteins. An extreme version of this is the use of docking to align molecules for CoMFA. Target-based methods such as FlexX are moving in a ligand-based direction, and CoMFA and CoMSIA (ligand-based methods) are moving towards a target-based approach.

The two approaches have also been combined by parallel application and in series; the former is a form of consensus scoring whereas the latter has been called "consensus screening." Scoring is used with surface feature complementarity in eHiTS [89]. Generation of a Structural Interaction Fingerprint (SIFt) [230] translates 3D structural binding information from a protein–ligand complex into a one-dimensional binary string. Each fingerprint represents the "structural interaction profile" of the complex that can be used to organize, analyze, and visualize the rich amount of information encoded in ligand–receptor complexes and also to assist database mining. Muegge and Oloff [231] have also discussed synergies between structure-based and ligand-based virtual screening.

14. Data Reduction and Visualization

Computer-aided drug design produces huge volumes of data, which are often multidimensional in nature. There is, thus, a need for methods to reduce the dimensionality of these data and visualize the results [232, 233]. Visualization techniques include self-organizing maps [44, 234, 235], tree maps [236–238], enhanced SAR Maps [44], dendrograms [239], radial clustergrams [240], non-linear maps [241–243], heatmaps [237], and various forms of conventional statistical plots (scatter plots, bar charts, pie charts, etc.). SAR trees [244] and scaffold trees [245, 246] make use of common substructures.

A SAR tree represents a collection of compounds as an acyclic graph, where each node represents a common substructure and its children represent the R-groups around it. Each R-group in turn embodies another common substructure that is shared by multiple compounds at that particular attachment site, and is recursively split into more refined R-groups, until there are no further variations. This method is also used in the commercially available ClassPharmer software [178].

Rule-based methods such as that of Bemis and Murcko [247] scale linearly with the number of structures since the classification

process is done individually for each molecule and incremental update is possible. The classes created by such methods are more intuitive to chemists than those produced by clustering and other methods. Chemical Abstracts Service has expanded on Bemis and Murcko's frameworks technique in developing SubScape, a new substance analysis and visualization tool for substances retrieved in SciFinder [248]. The SubScape Framework Identifier combines three identification numbers, each of which signifies one aspect of framework structure: a graph id denotes the underlying connectivity, a node id denotes the pattern of elements, and a bond id denotes the pattern of bond types.

The scaffold tree technique reported by Schuffenhauer and co-workers is also a variation on Bemis and Murcko's molecular frameworks. This hierarchical classification method [246] uses molecular frameworks as the leaf nodes of a scaffold tree. By iterative removal of rings, scaffolds forming the higher levels in the hierarchy tree are obtained. Prioritization rules ensure that less characteristic, peripheral rings are removed first. Highlighting by color intensity is used to show the fraction of active compounds containing a scaffold: this immediately identifies those branches of the scaffold tree which contain active molecules. Schuffenhauer and his colleagues [249] have compared rule-based and scaffold-oriented classification methods (in a Pareto analysis), and clustering based on molecular descriptors: no technique was found to be generally superior but all gave results which were to some extent biologically meaningful.

Agrafiotis et al. [250] have developed SAR maps to allow medicinal chemists to visualize structure-activity relationships. An SAR map renders an R-group decomposition of a congeneric series as a rectangular matrix of cells, each representing a unique combination of R-groups, color-coded by a property of the corresponding compound. An enhanced version of the software [251] expands the types of visualizations that can be displayed inside the cells. Examples include multidimensional histograms and pie charts that visualize the biological profiles of compounds across a panel of assays, forms that display specific fields on user-defined layouts, aligned 3D structure drawings that show the relative orientation of different substituents, dose–response curves, and images of crystals or diffraction patterns.

Medicinal chemists at ArQule can use the Spiral View tool [252] developed by Smellie. The relationship depiction paradigm of this tool is a spiral view centered on the most active compound, with the compounds most similar to it oriented clockwise around it. The chemist "walks around" the spiral. The width of the line between two compounds is proportional to the difference in property values between them. When the user clicks on a molecule, it then becomes the central molecule in a new spiral.

15. Text Mining

The discussion so far has covered many data mining techniques but drug-related information is also buried within written resources. Text mining is a relatively new technique used to extract this information automatically [253–258]. The first stage in this is information retrieval (IR) from the scientific literature, or patents, or in-house documents, which is carried out with general search engines or more specialized IR tools. The next step is information extraction (IE) from the text retrieved. Approaches to IE include rule-based or knowledge-based methods and statistical or machine-learning based methods. Named entity extraction (NER) recognizes terms for genes, proteins, drugs etc. NER can be based on dictionaries, rules, training sets (in machine learning) and combinations of these approaches. NER may involve simpler techniques such as co-occurrence of terms in the text, or the more sophisticated techniques of natural language processing.

In a next step, these extracted chemical names are converted into connection tables by means of a commercially available program (ACD/Labs, ChemAxon, CambridgeSoft, OpenEye Scientific Software, and InfoChem supply such software) and may be stored and retrieved in a chemical structure database system. A number of commercial organizations (e.g., TEMIS, Linguamatics, Notiora, IBM, SureChem, and InfoChem) have developed algorithms for chemical named entity recognition and have established relationships with cheminformatics companies such as Symyx, Accelrys, ChemAxon, and InfoChem. The publishing group of the Royal Society of Chemistry (RSC) has used the NER software Open Source Chemical Analysis Routines (OSCAR, developed at the University of Cambridge) to enhance articles with "live" chemical structures, in its system RSC Prospect [259–262]. RSC has used selections from the Open Biomedical Ontologies and it makes its own chemical ontologies freely available [263].

A more complex problem is the analysis of structure images, or "chemical optical character recognition": recognizing and distinguishing different graphical objects in a picture (structures, arrows, and text), and the conversion of the structure drawings into connection tables. To this end, four programs are currently under active development: Chemical Literature Date Extraction, CLiDE [264], chemoCR [265], Optical Structure Recognition Software, OSRA [266], and ChemReader [267]. Analyzing the image of a Markush structure is a challenge for the future: in this case there is the additional problem of finding the appropriate R-groups from the part of the text of the article or patent which acts as a key to the diagram in question.

16. Conclusion

Research continues into other applications involving Markush structures. For example, cheminformatics programs could be developed to include "freedom to operate" patent space in optimization studies. Text mining is still in its infancy: more chemical dictionaries and ontologies are constantly being developed and the performance of image recognition programs is gradually improving. At least three teams, at Key Module [268], Molecular Networks [39], and InfoChem [269] are active in the field of retrosynthesis. In computer-aided drug design, there is likely to be keen interest in machine learning in future: random forest seems to be gaining in popularity.

Matthews [270] has suggested many unmet needs in QSARs and expert systems, including integrated fragment and descriptor paradigms and 3D descriptors; QSARs based upon pure active ingredient and metabolites; QSARs for drug–drug interaction, for animal organ toxicities, and for regulatory dose concentration endpoints (e.g. lowest observed effect level and no observed effect level); and expert system rules for toxicities of substances such as biologicals which cannot be predicted by QSAR. Other unmet needs are databases of pharmaceutical off-target activities of pharmaceutical investigational new drugs, of confidential business information, and of regulatory dose concentration endpoints; integration of FDA and Environmental Protection Agency archival data; and advanced linguistic software to extract data.

Some of the challenges of structure-based drug design (docking and de novo design) have already been discussed. Prediction of affinity is a very hard problem and is as yet unsolved; Jorgensen believes that free energy guided molecular design, for example, may become a mainstream activity [131]. Ligands forming covalent complexes have been little studied. Proteins have been the major targets of docking methods. However, nucleic acids are also targets for medicinal chemistry and should be further investigated. Docking DNA intercalators is even more challenging [128].

In fragment-based drug design there are still many opportunities for new development, such as improving the novelty, structural diversity and physicochemical properties of fragment libraries, and improving detection methods to find hits with activity as low as 1–10 mM. It should be possible to identify new types of interactions: protein–protein interactions, novel templates, and new binding modes. There is much work to be done on increasing the efficiency of fragment optimization. One challenge is deciding which fragments to progress, other than using the subjective decisions of a medicinal chemist. Tools for assessing synthetic

accessibility may help. Optimizing fragments in the absence of a crystal structure is another hurdle. Progress in structural biology will lead to progress in FBDD.

Target-based and ligand-based methods have shared challenges. Targets are not rigid: account must be taken of alternative binding sites, tautomeric ambiguity, and accommodation; local and global binding site plasticity; and complex librational freedom. Ligands move too: the problem of discrete or complete conformational sampling is solved but tautomerism and pK_a are unsolved problems. Solvation and (de)solvation of ligands and targets is another shared challenge. Above all there is the enormous problem of entropy [271].

Practitioners of bioinformatics and cheminformatics have much to learn from each other. Oprea and colleagues have coined the term "systems chemical biology," believing that the future of cheminformatics lies in its ability to provide an integrative, predictive framework that links biological sciences [104]. Others are publishing in that field [272]. Computational chemical biology [273] certainly has a bright future.

References

1. Warr, W. A. Cheminformatics education. http://www.qsarworld.com/cheminformatics-education.php (accessed October 2, 2009).
2. Willett, P. (2008) A bibliometric analysis of the literature of chemoinformatics. *Aslib. Proc.* **60**, 4–17.
3. Cheminformatics or chemoinformatics? http://www.molinspiration.com/chemoinformatics.html (accessed October 2, 2009).
4. Warr, W. A. (1999) Balancing the needs of the recruiters and the aims of the educators, in *Book of Abstracts, 218th ACS National Meeting, New Orleans, Aug. 22–26.*
5. Warr, W. A. Extract from 218th ACS National Meeting and Exposition, New Orleans, Louisiana, August 1999 [cheminformatics]. http://www.warr.com/warrzone2000.html (accessed October 2, 2009).
6. Willett, P. (2007) A bibliometric analysis of the Journal of Molecular Graphics and Modelling. *J. Mol. Graphics Modell.* **26**, 602–606.
7. Leach, A. R., and Gillet, V. J. (2003) *An Introduction to Chemoinformatics.* Kluwer, Dordrecht, The Netherlands.
8. Gasteiger, J., and Engel, T., (Eds.) (2003) *Chemoinformatics: A Textbook*, Wiley-VCH, Weinheim, Germany.
9. Bajorath, J., (Ed.) (2004) *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*, Humana Press, Totowa, NJ.
10. Oprea, T. I., (Ed.) (2005) *Chemoinformatics in Drug Discovery*, Wiley, New York, NY.
11. Schneider, G., and Baringhaus, K.-H., (Eds.) (2008) *Molecular Design: Concepts and Applications*, Wiley-VCH, Weinheim, Germany.
12. Chen, W. L. (2006) Chemoinformatics: past, present, and future. *J. Chem. Inf. Model.* **46**, 2230–2255.
13. Engel, T. (2006) Basic overview of chemoinformatics. *J. Chem. Inf. Model.* **46**, 2267–2277.
14. Willett, P. (2008) From chemical documentation to chemoinformatics: 50 years of chemical information science. *J. Inf. Sci.* **34**, 477–499.
15. Willett, P. (2003) A history of chemoinformatics in *Handbook of Chemoinformatics: From Data to Knowledge* Wiley-VCH, Weinheim, Germany, Vol. 1, pp 6–20.
16. Bishop, N., Gillet, V. J., Holliday, J. D., and Willett, P. (2003) Chemoinformatics research at the University of Sheffield: a history and citation analysis. *J. Inf. Sci.* **29**, 249–267.
17. Ash, J. E., Warr, W. A., and Willett, P., (Eds.) (1991) *Chemical structure systems: computational techniques for representation, searching, and processing of structural information*, Ellis Horwood, Chichester, UK.

18. Paris, G. C. (1997) Chemical structure handling by computer. *Ann. Rev. Inf. Sci. Technol.* **32**, 271–337.
19. Paris, G. C. (1998) Structure databases, in *Encyclopedia of Computational Chemistry* (Schleyer, P. v. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, and Schreiner, P. R., Eds.), Wiley, Chichester, UK, Vol. 4, pp 2771–2785.
20. Paris, G. C. (2003) Databases of chemical structures, in *Handbook of Cheminformatics: From Data to Knowledge* (Gasteiger, J., Ed.), Wiley-VCH, Weinheim, Germany, Vol. 2, pp 523–555.
21. Warr, W. A., and Suhr, C. (1992) *Chemical information management*, Wiley-VCH, Weinheim, Germany.
22. Ott, M., A. (2004) Cheminformatics and organic chemistry. Computer-assisted synthetic analysis, in *Cheminformatics Developments* (Noordik, J., H., Ed.), IOS Press, Amsterdam, The Netherlands, pp 83–109.
23. Pearlman, R. S. (1987) Rapid generation of high quality approximate 3D molecular structures. *Chem. Des. Autom. News* **2**, 5–7.
24. Gasteiger, J., Rudolph, C., and Sadowski, J. (1990) Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **3**, 537–547.
25. Hiller, C., and Gasteiger, J. (1987) An automatic molecule builder, in *Software Development in Chemistry 1. Proceedings of the Workshops on the Computer in Chemistry, Hochfilzen/Tirol, November 19–21, 1986* (Gasteiger, J., Ed.), Springer, Berlin, Vol. 1, pp 53–66.
26. Sadowski, J., Gasteiger, J., and Klebe, G. (1994) Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **34**, 1000–1008.
27. Jakes, S. E., and Willett, P. (1986) Pharmacophoric pattern matching in files of 3-D chemical structures: selection of interatomic distance screens. *J. Mol. Graphics* **4**, 12–20.
28. Jakes, S. E., Watts, N., Willett, P., Bawden, D., and Fisher, J. D. (1987) Pharmacophoric pattern matching in files of 3D chemical structures: evaluation of search performance. *J. Mol. Graphics* **5**, 41–48.
29. Brint, A. T., and Willett, P. (1987) Pharmacophoric pattern matching in files of 3D chemical structures: comparison of geometric searching algorithms. *J. Mol. Graphics* **5**, 49–56.
30. Cringean, J. K., Pepperrell, C. A., Poirrette, A. R., and Willett, P. (1990) Selection of screens for three-dimensional substructure searching. *Tetrahedron Comput. Methodol.* **3**, 37–46.
31. Willett, P. (1991) *Three-dimensional Chemical Structure Handling*, Research Studies Press, Taunton, UK.
32. Martin, Y. C., and Willett, P., (Eds.) (1998) *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*, American Chemical Society, Washington, DC.
33. Martin, Y. C., Danaher, E. B., May, C. S., and Weininger, D. (1988) MENTHOR, a database system for the storage and retrieval of three-dimensional molecular structures and associated data searchable by substructural, biologic, physical, or geometric properties. *J. Comput.-Aided Mol. Des.* **2**, 15–29.
34. Van Drie, J. H., Weininger, D., and Martin, Y. C. (1989) ALADDIN: an integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure searching of three-dimensional molecular structures. *J. Comput.-Aided Mol. Des.* **3**, 225–251.
35. Willett, P., Barnard, J. M., and Downs, G. M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996.
36. Barnard, J. M. (1991) A comparison of different approaches to Markush structure handling. *J. Chem. Inf. Comput. Sci.* **31**, 64–68.
37. Benichou, P., Klimczak, C., and Borne, P. (1997) Handling genericity in chemical structures using the Markush DARC software. *J. Chem. Inf. Comput. Sci.* **37**, 43–53.
38. Corey, E. J., and Wipke, W. T. (1969) Computer-assisted design of complex organic syntheses. *Science* **166**, 178–192.
39. THERESA. Molecular Networks. http://www.molecular-networks.com/products/theresa (accessed October 2, 2009).
40. Dugundji, J., and Ugi, I. (1973) Algebraic model of constitutional chemistry as a basis for chemical computer programs. *Fortschr. Chem. Forsch.* **39**, 19–64.
41. Bauer, J., Herges, R., Fontain, E., and Ugi, I. (1985) IGOR and computer assisted innovation in chemistry. *Chimia* **39**, 43–53.
42. Gasteiger, J., Ihlenfeldt, W. D., Roese, P., and Wanke, R. (1990) Computer-assisted reaction prediction and synthesis design. *Anal. Chim. Acta* **235**, 65–75.
43. Jorgensen, W. L., Laird, E. R., Gushurst, A. J., Fleischer, J. M., Gothe, S. A., Helson, H. E., Paderes, G. D., and Sinclair, S. (1990) CAMEO: a program for the logical prediction of the products of organic reactions. *Pure Appl. Chem.* **62**, 1921–1932.
44. Sadowski, J., Wagener, M., and Gasteiger, J. (1996) Assessing similarity and diversity

of combinatorial libraries by spatial autocorrelation functions and neural networks. *Angew. Chem. Int. Ed. Engl.* **34**, 2674–2677.
45. Sadowski, J., Gasteiger, J., and Klebe, G. (2002) Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inf. Comput. Sci.* **34**, 1000–1008.
46. Kennard, O., Watson, D. G., and Town, W. G. (1972) Cambridge Crystallographic Data Centre. I. Bibliographic file. *J. Chem. Doc.* **12**, 14–19.
47. Allen, F. H., Kennard, O., Motherwell, W. D. S., Town, W. G., and Watson, D. G. (1973) Cambridge Crystallographic Data Centre. II. Structural data file. *J. Chem. Doc.* **13**, 119–123.
48. Allen, F. H. (2002) The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr. Sect. B Struct. Sci.* **B58**, 380–388.
49. Allen, F. H., Battle, G., and Robertson, S. (2007) The Cambridge Structural Database, in *Comprehensive Medicinal Chemistry II.* (Triggle, D. J., and Taylor, J. B., Eds.), Elsevier, Amsterdam, The Netherlands, Vol. 3, pp 389–410.
50. Berman, H. M. (2008) The Protein Data Bank: a historical perspective. *Acta Crystallogr., Sect. A Found. Crystallogr.* **A64**, 88–95.
51. Martin, Y. C. (1998) Pharmacophore mapping, in *Designing Bioactive Molecules* (Martin, Y. C., and Willett, P., Eds.), American Chemical Society, Washington, DC, pp 121–148.
52. Martin, Y. C., Bures, M. G., Danaher, E. A., DeLazzer, J., Lico, I., and Pavlik, P. A. (1993) A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Des.* **7**, 83–102.
53. Greene, J., Kahn, S., Savoj, H., Sprague, P., and Teig, S. (1994) Chemical function queries for 3D database search. *J. Chem. Inf. Comput. Sci.* **34**, 1297–1308.
54. Sprague, P. (1995) Automated chemical hypothesis generation and database searching with Catalyst. *Perspect. Drug Discov. Des.* **3**, 1–20.
55. Jones, G., Willett, P., and Glen, R. C. (1995) A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **9**, 532–549.
56. Jones, G., Willett, P., and Glen, R. C. (1995) Molecular recognition of a receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **245**, 43–53.
57. Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. (1997) Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **267**, 727–748.
58. Winkler, D. A. (2002) The role of quantitative structure-activity relationships (QSAR) in biomolecular discovery. *Briefings Bioinf.* **3**, 73–86.
59. Tropsha, A. (2003) Recent trends in quantitative structure-activity relationships, in *Burger's Medicinal Chemistry and Drug Discovery, Volume 1, Drug Discovery* (Abraham, D. J., Ed.) 6th ed., Wiley, New York, NY.
60. Tropsha, A. (2005) Application of predictive QSAR models to database mining, in *Cheminformatics in Drug Discovery* (Oprea, T. I., Ed.), Wiley, New York, NY, pp 437–455.
61. Gramatica, P. A short history of QSAR evolution. http://www.qsarworld.com/Temp_Fileupload/Shorthistoryofqsar.pdf (accessed September 23, 2009).
62. Hawkins, D. M. QSAR approaches, models and statistics relating to toxicity prediction. In W. A Warr. Proceedings of New Horizons in Toxicity Prediction. Lhasa Limited symposium event in collaboration with the University of Cambridge, December 2008. http://www.qsarworld.com/files/Lhasa_Symposium_2008_Report.pdf (accessed September 23, 2009).
63. Cramer, R. D., Patterson, D. E., and Bunce, J. D. (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110**, 5959–5967.
64. Klebe, G., Abraham, U., and Mietzner, T. (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J. Med. Chem.* **37**, 4130–4146.
65. Lemmen, C., Lengauer, T., and Klebe, G. (1998) FlexS: a method for fast flexible ligand superposition. *J. Med. Chem.* **41**, 4502–4520.
66. Kennedy, T. (1997) Managing the drug discovery/development interface. *Drug Discov. Today* **2**, 436–444.
67. Benigni, R., (Ed.) (2003) *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*, CRC Press, Boca Raton, FL.
68. Cronin, M. T. D., and Livingstone, D. J., (Eds.) (2004) *Predicting Chemical Toxicity and Fate*, CRC Press, Boca Raton, FL.
69. Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery

and development settings. *Adv. Drug Deliv. Rev.* **23**, 3–25.
70. Cramer, R. (1995) Unpublished work.
71. Güner, O. F., (Ed.) (2000) *Pharmacophore: Perception, Development, and Use in Drug Design. [In: IUL Biotechnol. Ser., 2000; 2]*, International University Line, La Jolla, CA.
72. Willett, P., (Ed.) (1997) *Computational Methods for the Analysis of Molecular Diversity (Perspectives in Drug Discovery and Design 1997, Volumes 7/8)*, Kluwer/Escom, Dordrecht, The Netherlands.
73. Brown, R. D., and Martin, Y. C. (1996) Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **36**, 572–584.
74. Brown, R. D., and Martin, Y. C. (1997) The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **37**, 1–9.
75. Martin, E. J., Blaney, J. M., Siani, M. A., Spellmeyer, D. C., Wong, A. K., and Moos, W. H. (1995) Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* **38**, 1431–1436.
76. Gillet, V. J., Willett, P., and Bradshaw, J. (1997) The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **37**, 731–740.
77. Cottrell, S. J., Gillet, V. J., Taylor, R., and Wilton, D. J. (2004) Generation of multiple pharmacophore hypotheses using multiobjective optimisation techniques. *J. Comput.-Aided Mol. Des.* **18**, 665–682.
78. Cottrell, S., Gillet, V., and Taylor, R. (2006) Incorporating partial matches within multiobjective pharmacophore identification. *J. Comput.-Aided Mol. Des.* **20**, 735–749.
79. Teague, S., J., Davis, A., M., Leeson, P., D., and Oprea, T. I. (1999) The design of leadlike combinatorial libraries. *Angew. Chem. Int. Ed. Engl.* **38**, 3743–3748.
80. Oprea, T. I., Davis, A. M., Teague, S. J., and Leeson, P. D. (2001) Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **41**, 1308–1315.
81. Hann, M. M., Leach, A. R., and Harper, G. (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **41**, 856–864.
82. Sadowski, J., and Kubinyi, H. (1998) A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **41**, 3325–3329.
83. Wagener, M., and Van Geerestein, V. J. (2000) Potential drugs and nondrugs: prediction and identification of important structural features. *J. Chem. Inf. Comput. Sci.* **40**, 280–292.
84. Willett, P. (2006) Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **11**, 1046–1053.
85. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E. (1982) A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161**, 269–288.
86. Rarey, M., Kramer, B., Lengauer, T., and Klebe, G. (1996) A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **261**, 470–489.
87. McMeeking, B., and Fletcher, D. (2004) The United Kingdom Chemical Database Service: CDS, in *Cheminformatics Developments* (Noordik, J., H., Ed.), IOS Press, Amsterdam, The Netherlands, pp 37–67.
88. (Multiple authors) (2006) Focus on Cyberinfrastructure ("e-science"), the Enabling Platform for Cheminformatics. *J. Chem. Inf. Model.* **46**.
89. Zsoldos, Z., Reid, D., Simon, A., Sadjad, S. B., and Johnson, A. P. (2007) eHiTS: a new fast, exhaustive flexible ligand docking system. *J. Mol. Graphics Modell.* **26**, 198–212.
90. Salamone, S. GTA4: enabler of life sciences research? http://www.bio-itworld.com/inside-it/2008/2005/gta2004-and-life-sciences.html (accessed October 2, 2009).
91. Murray-Rust, P. (2008) Chemistry for everyone. *Nature (London, U. K.)* **451**, 648–651.
92. Williams, A. J. (2008) Internet-based tools for communication and collaboration in chemistry. *Drug Discov. Today* **13**, 502–506.
93. Warr, W. A. (2008) Social software: fun and games, or business tools? *J. Inf. Sci.* **34**, 591–604.
94. Coles, S. J., Day, N. E., Murray-Rust, P., Rzepa, H. S., and Zhang, Y. (2005) Enhancement of the chemical semantic web through the use of InChI identifiers. *Org. Biomol. Chem.* **3**, 1832–1834.
95. Guha, R., Howard, M. T., Hutchison, G. R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J., and Willighagen, E. L. (2006) The Blue Obelisk. Interoperability in chemical informatics. *J. Chem. Inf. Model.* **46**, 991–998.
96. Taylor, K. R., Gledhill, R. J., Essex, J. W., Frey, J. G., Harris, S. W., and De Roure, D. C.

(2006) Bringing chemical data onto the semantic web. *J. Chem. Inf. Model.* **46**, 939–952.
97. Frey, J. G. (2009) The value of the Semantic Web in the laboratory. *Drug Discov. Today* **14**, 552–561.
98. Gardner, B. Approaches to information integration. Paper given at ICIC 2007, Sitges, Spain, October 21–24, 2007. http://www.infonortics.eu/chemical/ch07/slides/gardner.pdf (accessed September 18, 2009).
99. Symyx Technologies. DiscoveryGate. http://www.discoverygate.com (accessed October 2, 2009).
100. ChemSpider. http://www.chemspider.com (accessed September 18, 2009).
101. Warr, W. A. Workflow and pipelining in cheminformatics. http://www.qsarworld.com/qsar-workflow1.php (accessed October 2, 2009).
102. Jorgensen, W. L. (2004) The many roles of computation in drug discovery. *Science* **303**, 1813–1818.
103. PubChem. http://pubchem.ncbi.nlm.nih.gov/search/search.cgi (accessed September 18, 2009).
104. Oprea, T. I., Tropsha, A., Faulon, J.-L., and Rintoul, M. D. (2007) Systems chemical biology. *Nat. Chem. Biol.* **3**, 447–450.
105. Therapeutics for Rare and Neglected Diseases. http://www.nih.gov/news/health/may2009/nhgri-2020.htm (accessed October 7, 2009).
106. DeLano, W. L. (2005) The case for open-source software in drug discovery. *Drug Discov. Today* **10**, 213–217.
107. Stahl, M. T. (2005) Open-source software: not quite endsville. *Drug Discov. Today* **10**, 219–222.
108. The Pistoia Alliance. http://pistoiaalliance.org/ (accessed October 7, 2009).
109. Barnes, M. R., Harland, L., Foord, S. M., Hall, M. D., Dix, I., Thomas, S., Williams-Jones, B. I., and Brouwer, C. R. (2009) Lowering industry firewalls: pre-competitive informatics initiatives in drug discovery. *Nat. Rev. Drug Discov.* **8**, 701–708.
110. Lhasa Limited. Derek for Windows. http://www.lhasalimited.org/ (accessed October 7, 2009).
111. Irwin, J. J., and Shoichet, B. K. (2004) ZINC – a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182.
112. eMolecules. http://www.emolecules.com (accessed September 18, 2009).
113. NIH/CADD chemical structure lookup service. http://cactus.nci.nih.gov/lookup (accessed October 2, 2009).
114. Williams, A. J. (2008) A perspective of publicly accessible/open-access chemistry databases. *Drug Discov. Today* **13**, 495–501.
115. Fink, T., and Reymond, J.-L. (2007) Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **47**, 342–353.
116. Blum, L. C., and Reymond, J.-L. (2009) 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* **131**, 8732–8733.
117. Collaborative Drug Discovery. http://www.collaborativedrug.com/ (accessed September 18, 2009).
118. Morgan, H. L. (1965) The generation of a unique machine description for chemical structures – a technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **5**, 107–113.
119. Wipke, W. T., and Dyott, T. M. (1974) Stereochemically unique naming algorithm. *J. Am. Chem. Soc.* **96**, 4834–4842.
120. Hillard, R., and Taylor, K. T. (2009) InChI keys as standard global identifiers in chemistry web services, in *Abstracts of Papers, 237th ACS National Meeting, Salt Lake City, UT, United States, March 22–26, 2009.*
121. The IUPAC International Chemical Identifier (InChI). http://www.iupac.org/inchi/ (accessed September 18, 2009).
122. Heller, S. R., and McNaught, A. D. (2009) The IUPAC international chemical identifier (InChI). *Chem. Int.* **31**, 7–9.
123. Warr, W. A. The IUPAC International Chemical Identifier. http://www.qsarworld.com/INCHI1.php (accessed September 18, 2009).
124. McKay, B. D. (1981) Practical graph isomorphism. *Congressus Numeratium* **30**, 45–87.
125. Weininger, D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36.
126. Weininger, D., Weininger, A., and Weininger, J. L. (1989) SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **29**, 97–101.
127. Leach, A. R., Shoichet, B. K., and Peishoff, C. E. (2006) Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **49**, 5851–5855.

128. Moitessier, N., Englebienne, P., Lee, D., Lawandi, J., and Corbeil, C. R. (2008) Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **153**, S7-S26.
129. Wang, R., Lai, L., and Wang, S. (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **16**, 11–26.
130. Waszkowycz, B. (2008) Towards improving compound selection in structure-based virtual screening. *Drug Discov. Today* **13**, 219–226.
131. Jorgensen, W. L. (2009) Efficient drug lead discovery and optimization. *Acc. Chem. Res.* **42**, 724–733.
132. Irwin, J. (2008) Community benchmarks for virtual screening. *J. Comput.-Aided Mol. Des.* **22**, 193–199.
133. Liebeschuetz, J. (2008) Evaluating docking programs: keeping the playing field level. *J. Comput.-Aided Mol. Des.* **22**, 229–238.
134. DUD. A Directory of Useful Decoys. http://dud.docking.org/ (accessed September 18, 2009).
135. Huang, N., Shoichet, B. K., and Irwin, J. J. (2006) Benchmarking sets for molecular docking. *J. Med. Chem.* **49**, 6789–6801.
136. Good, A., and Oprea, T. (2008) Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J. Comput.-Aided Mol. Des.* **22**, 169–178.
137. Jain, A. (2008) Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput.-Aided Mol. Des.* **22**, 201–212.
138. Jain, A., and Nicholls, A. (2008) Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **22**, 133–139.
139. Cole, J. C., Murray, C. W., Nissink, J. W. M., Taylor, R. D., and Taylor, R. (2005) Comparing protein-ligand docking programs is difficult. *Proteins Struct., Funct., Bioinf.* **60**, 325–332.
140. Clark, R., and Webster-Clark, D. (2008) Managing bias in ROC curves. *J. Comput.-Aided Mol. Des.* **22**, 141–146.
141. Nicholls, A. (2008) What do we know and when do we know it? *J. Comput.-Aided Mol. Des.* **22**, 239–255.
142. Truchon, J.-F., and Bayly, C. I. (2007) Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **47**, 488–508.
143. Sheridan, R. P., Singh, S. B., Fluder, E. M., and Kearsley, S. K. (2001) Protocols for bridging the peptide to nonpeptide gap in topological similarity searches. *J. Chem. Inf. Comput. Sci.* **41**, 1395–1406.
144. Hawkins, P. C. D., Skillman, A. G., and Nicholls, A. (2006) Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **50**, 74–82.
145. Zhang, Q., and Muegge, I. (2006) Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting, and consensus scoring. *J. Med. Chem.* **49**, 1536–1548.
146. McGaughey, G. B., Sheridan, R. P., Bayly, C. I., Culberson, J. C., Kreatsoulas, C., Lindsley, S., Maiorov, V., Truchon, J.-F., and Cornell, W. D. (2007) Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **47**, 1504–1519.
147. Rush, T. S., Grant, J. A., Mosyak, L., and Nicholls, A. (2005) A shape-based 3-D scaffold hopping method and its application to a bacterial protein–protein interaction. *J. Med. Chem.* **48**, 1489–1495.
148. Verdonk, M. L., Berdini, V., Hartshorn, M. J., Mooij, W. T. M., Murray, C. W., Taylor, R. D., and Watson, P. (2004) Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **44**, 793–806.
149. Klebe, G. (2006) Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today* **11**, 580–594.
150. Böhm, H.-J. (1992) The computer program LUDI: a new method for the *de novo* design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* **6**, 61–78.
151. Law, J. M. S., Fung, D. Y. K., Zsoldos, Z., Simon, A., Szabo, Z., Csizmadia, I. G., and Johnson, A. P. (2003) Validation of the SPROUT *de novo* design program. *Theochem* **666–667**, 651–657.
152. Boda, K., and Johnson, A. P. (2006) Molecular complexity analysis of *de novo* designed ligands. *J. Med. Chem.* **49**, 5869–5879.
153. Barreiro, G., Kim, J. T., Guimaraes, C. R. W., Bailey, C. M., Domaoal, R. A., Wang, L., Anderson, K. S., and Jorgensen, W. L. (2007) From docking false-positive to active anti-HIV agent. *J. Med. Chem.* **50**, 5324–5329.
154. Baber, J. C., and Feher, M. (2004) Predicting synthetic accessibility: application in drug discovery and development. *Mini-Rev. Med. Chem.* **4**, 681–692.
155. Boda, K., Seidel, T., and Gasteiger, J. (2007) Structure and reaction based evaluation of

synthetic accessibility. *J. Comput.-Aided Mol. Des.* **21**, 311–325.
156. Zaliani, A., Boda, K., Seidel, T., Herwig, A., Schwab, C. H., Gasteiger, J., Claussen, H., Lemmen, C., Degen, J., Paern, J., and Rarey, M. (2009) Second-generation de novo design: a view from a medicinal chemist perspective. *J. Comput.-Aided Mol. Des.* **23**, 593–602.
157. Patel, H., Bodkin, M. J., Chen, B., and Gillet, V. J. (2009) Knowledge-based approach to *de novo* design using reaction vectors. *J. Chem. Inf. Model.* **49**, 1163–1184.
158. Shuker, S. B., Hajduk, P. J., Meadows, R. P., and Fesik, S. W. (1996) Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **274**, 1531–1534.
159. Allen, K. N., Bellamacina, C. R., Ding, X., Jeffery, C. J., Mattos, C., Petsko, G. A., and Ringe, D. (1996) An experimental approach to mapping the binding surfaces of crystalline proteins. *J. Phys. Chem.* **100**, 2605–2611.
160. Carr, R. A. E., Congreve, M., Murray, C. W., and Rees, D. C. (2005) Fragment-based lead discovery: leads by design. *Drug Discov. Today* **10**, 987–992.
161. Warr, W. (2009) Fragment-based drug discovery. *J. Comput.-Aided Mol. Des.* **23**, 453-458.
162. Joseph-McCarthy, D. (2009) Challenges of fragment screening. *J. Comput.-Aided Mol. Des.* **23**, 449–451.
163. Chen, I. J., and Hubbard, R. (2009) Lessons for fragment library design: analysis of output from multiple screening campaigns. *J. Comput.-Aided Mol. Des.* **23**, 603–620.
164. Hopkins, A. L., Groom, C. R., and Alex, A. (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **9**, 430–431.
165. Abad-Zapatero, C., and Metz James, T. (2005) Ligand efficiency indices as guideposts for drug discovery. *Drug Discov. Today* **10**, 464–469.
166. Leeson, P. D., and Springthorpe, B. (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discov.* **6**, 881–890.
167. Congreve, M., Chessari, G., Tisi, D., and Woodhead, A. J. (2008) Recent developments in fragment-based drug discovery. *J. Med. Chem.* **51**, 3661–3680.
168. Congreve, M., Carr, R., Murray, C., and Jhoti, H. (2003) A "rule of three" for fragment-based lead discovery? *Drug Discov. Today* **8**, 876–877.
169. Hajduk, P. J., and Greer, J. (2007) A decade of fragment-based drug design: strategic advances and lessons learned. *Nat. Rev. Drug Discov.* **6**, 211–219.
170. Erlanson, D. A., Wells, J. A., and Braisted, A. C. (2004) Tethering: fragment-based drug discovery. *Annu. Rev. Biophys. Biomol. Struct.* **33**, 199–223, 194 plates.
171. Blomberg, N., Cosgrove, D., Kenny, P., and Kolmodin, K. (2009) Design of compound libraries for fragment screening. *J. Comput.-Aided Mol. Des.* **23**, 513–525.
172. de Kloe, G. E., Bailey, D., Leurs, R., and de Esch, I. J. P. (2009) Transforming fragments into candidates: small becomes big in medicinal chemistry. *Drug Discov. Today* **14**, 630–646.
173. Willett, P. (2009) Similarity methods in chemoinformatics. *Ann. Rev. Inf. Sci. Technol.* **43**, 3–71.
174. Eckert, H., and Bajorath, J. (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today* **12**, 225–233.
175. Raymond, J. W., and Willett, P. (2002) Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **16**, 521–533.
176. Ghose, A. K., Herbertz, T., Salvino, J. M., and Mallamo, J. P. (2006) Knowledge-based chemoinformatic approaches to drug discovery. *Drug Discov. Today* **11**, 1107–1114.
177. Accelrys. Pipeline Pilot http://accelrys.com/products/scitegic/ (accessed October 2, 2009).
178. Simulations Plus. ClassPharmer. http://www.simulations-plus.com/ (accessed October 6, 2009).
179. ChemAxon. Library MCS. http://www.chemaxon.com/shared/libMCS/ (accessed October 2, 2009).
180. Cramer, R. D. (2003) Topomer CoMFA: a design methodology for rapid lead optimization. *J. Med. Chem.* **46**, 374–388.
181. Jilek, R. J., and Cramer, R. D. (2004) Topomers: a validated protocol for their self-consistent generation. *J. Chem. Inf. Comput. Sci.* **44**, 1221–1227.
182. Cramer, R., and Wendt, B. (2007) Pushing the boundaries of 3D-QSAR. *J. Comput.-Aided Mol. Des.* **21**, 23–32.
183. Lill, M. A. (2007) Multi-dimensional QSAR in drug discovery. *Drug Discov. Today* **12**, 1013–1017.
184. Hopfinger, A. J., Wang, S., Tokarski, J. S., Jin, B., Albuquerque, M., Madhav, P. J., and Duraiswami, C. (1997) Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **119**, 10509–10524.

185. Stouch, T. R., Kenyon, J. R., Johnson, S. R., Chen, X.-Q., Doweyko, A., and Li, Y. (2003) *In silico* ADME/Tox: why models fail. *J. Comput.-Aided Mol. Des.* **17**, 83–92.
186. Kubinyi, H. Why models fail. http://americanchemicalsociety.mediasite.com/acs/viewer/?peid=7a194d147-baa199-19-4b192d-a823-191fd117bf5301c (accessed September 22, 2009).
187. Tropsha, A., and Golbraikh, A. (2007) Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.* **13**, 3494–3504.
188. Hawkins, D. M. (2003) The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **44**, 1–12.
189. Sheridan, R. P., Feuston, B. P., Maiorov, V. N., and Kearsley, S. K. (2004) Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **44**, 1912–1928.
190. Kubinyi, H., Hamprecht, F. A., and Mietzner, T. (1998) Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. *J. Med. Chem.* **41**, 2553–2564.
191. Kubinyi, H. (2006) Validation and predictivity of QSAR models, in *QSAR & Molecular Modelling in Rational Design of Bioactive Molecules, Proceedings of the 15th European Symposium on Structure-Activity Relationships and Molecular Modelling, Istanbul, Turkey 2004*, pp 30–33.
192. Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.-D., Lee, K.-H., and Tropsha, A. (2003) Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des.* **17**, 241–253.
193. Golbraikh, A., and Tropsha, A. (2002) Beware of q2! *J. Mol. Graphics Modell.* **20**, 269–276.
194. Doweyko, A. M. (2004) 3D-QSAR illusions. *J. Comput.-Aided Mol. Des.* **18**, 587–596.
195. Gramatica, P. (2007) Principles of QSAR models validation: internal and external. *QSAR Comb. Sci.* **26**, 694–701.
196. Tropsha, A., Gramatica, P., and Gombar, V. K. (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR & Comb. Sci.* **22**, 69–77.
197. Jorgensen, W. L. (2006) QSAR/QSPR and proprietary data. *J. Chem. Inf. Model.* **46**, 937–937.
198. Editorial. (2006) QSAR/QSPR and proprietary data. *J. Med. Chem.* **49**, 3431–3431.
199. *ChemMedChem* notice to authors. http://www3.interscience.wiley.com/journal/110485305/home/110482452_notice.html?CRETRY=110485301&SRETRY=110485300 (accessed September 22, 2009).
200. Maggiora, G. M. (2006) On outliers and activity cliffs. Why QSAR often disappoints. *J. Chem. Inf. Model.* **46**, 1535–1535.
201. Peltason, L., and Bajorath, J. (2007) Molecular similarity analysis uncovers heterogeneous structure-activity relationships and variable activity landscapes. *Chem. Biol.* **14**, 489–497.
202. Peltason, L., and Bajorath, J. (2007) SAR Index: quantifying the nature of structure-activity relationships. *J. Med. Chem.* **50**, 5571–5578.
203. Guha, R., and Van Drie, J. H. (2008) Assessing how well a modeling protocol captures a structure-activity landscape. *J. Chem. Inf. Model.* **48**, 1716–1728.
204. Guha, R., and Van Drie, J. H. (2008) Structure-activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **48**, 646–658.
205. Medina-Franco, J. L., Martinez-Mayorga, K., Bender, A., Marin, R. M., Giulianotti, M. A., Pinilla, C., and Houghten, R. A. (2009) Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. *J. Chem. Inf. Model.* **49**, 477–491.
206. Johnson, S. R. (2007) The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model.* **48**, 25–26.
207. Pavan, M., and Worth, A. P. Review of QSAR Models for Ready Biodegradation. http://ecb.jrc.ec.europa.eu/documents/QSAR/QSAR_Review_Biodegradation.pdf (accessed September 23, 2009).
208. Warr, W. A. Proceedings of New Horizons in Toxicity Prediction. Lhasa Limited symposium event in collaboration with the University of Cambridge, December 2008. http://www.qsarworld.com/files/Lhasa_Symposium_2008_Report.pdf (accessed September 23, 2009).
209. Huynh, L., Masereeuw, R., Friedberg, T., Ingelman-Sundberg, M., and Manivet, P. (2009) In silico platform for xenobiotics ADME-T pharmacological properties modeling and prediction. Part I: beyond the reduction of animal model use. *Drug Discov. Today* **14**, 401–405.
210. Gramatica, P., Pilutti, P., and Papa, E. (2004) Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting

into training-test sets and consensus modeling. *J. Chem. Inf. Comput. Sci.* **44**, 1794–1802.
211. Hewitt, M., Cronin, M. T. D., Madden, J. C., Rowe, P. H., Johnson, C., Obi, A., and Enoch, S. J. (2007) Consensus QSAR models: do the benefits outweigh the complexity? *J. Chem. Inf. Model.* **47**, 1460–1468.
212. Matthews, E. J., Kruhlak, N. L., Benz, R. D., Contrera, J. F., Marchant, C. A., and Yang, C. (2008) Combined use of MC4PC, MDL-QSAR, BioEpisteme, Leadscope PDM, and Derek for Windows software to achieve high-performance, high-confidence, mode of action-based predictions of chemical carcinogenesis in rodents. *Toxicol. Mech. Methods* **18**, 189–206.
213. Abshear, T., Banik, G. M., D'Souza, M. L., Nedwed, K., and Peng, C. (2006) A model validation and consensus building environment. *SAR QSAR Environ. Res.* **17**, 311–321.
214. Boyer, S., Arnby, C. H., Carlsson, L., Smith, J., Stein, V., and Glen, R. C. (2007) Reaction site mapping of xenobiotic biotransformations. *J. Chem. Inf. Model.* **47**, 583–590.
215. Terfloth, L., Bienfait, B., and Gasteiger, J. (2007) Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. *J. Chem. Inf. Model.* **47**, 1688–1701.
216. Molecular Networks. BioPath. http://www.mol-net.de/biopath/index.html (accessed September 25, 2009).
217. Schwaighofer, A., Schroeter, T., Mika, S., Hansen, K., ter Laak, A., Lienau, P., Reichel, A., Heinrich, N., and Müller, K.-R. (2008) A probabilistic approach to classifying metabolic stability. *J. Chem. Inf. Model.* **48**, 785–796.
218. Symyx Technologies. Metabolite. http://www.symyx.com/products/databases/bioactivity/metabolite/index.jsp (accessed October 2, 2009).
219. Ridder, L., and Wagener, M. (2008) SyGMa: combining expert knowledge and empirical scoring in the prediction of metabolites. *ChemMedChem* **3**, 821–832.
220. Button, W. G., Judson, P. N., Long, A., and Vessey, J. D. (2003) Using absolute and relative reasoning in the prediction of the potential metabolism of xenobiotics. *J. Chem. Inf. Comput. Sci.* **43**, 1371–1377.
221. Leach, A. R., Gillet, V. J., Lewis, R. A., and Taylor, R. (2010) Three-Dimensional Pharmacophore Methods in Drug Discovery. J. Med. Chem. **53**, 539–558
222. Wolber, G., Seidel, T., Bendix, F., and Langer, T. (2008) Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discov. Today* **13**, 23–29.
223. Feng, J., Sanil, A., and Young, S. S. (2006) PharmID: pharmacophore identification using Gibbs sampling. *J. Chem. Inf. Model.* **46**, 1352–1359.
224. Richmond, N. J., Willett, P., and Clark, R. D. (2004) Alignment of three-dimensional molecules using an image recognition algorithm. *J. Mol. Graphics Modell.* **23**, 199–209.
225. Richmond, N., Abrams, C., Wolohan, P., Abrahamian, E., Willett, P., and Clark, R. (2006) GALAHAD: 1. Pharmacophore identification by hypermolecular alignment of ligands in 3D. *J. Comput.-Aided Mol. Des.* **20**, 567–587.
226. Jain, A. N. (2004) Ligand-based structural hypotheses for virtual screening. *J. Med. Chem.* **47**, 947–961.
227. Cole, J. C., Gardiner, E. J., Gillet, V. J., and Taylor, R. (2009) Development of test systems for pharmacophore elucidation, in *Abstracts of Papers, 237th ACS National Meeting, Salt Lake City, UT, United States, March 22–26, 2009.*
228. Patel, Y., Gillet, V. J., Bravi, G., and Leach, A. R. (2002) A comparison of the pharmacophore identification programs: catalyst, DISCO and GASP. *J. Comput.-Aided Mol. Des.* **16**, 653–681.
229. Clark, R. D. (2009) Prospective ligand- and target-based 3D QSAR: state of the art 2008. *Curr. Top. Med. Chem.* **9**, 791–810.
230. Deng, Z., Chuaqui, C., and Singh, J. (2003) Structural Interaction Fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding Interactions. *J. Med. Chem.* **47**, 337–344.
231. Muegge, I., and Oloff, S. (2006) Advances in virtual screening. *Drug Discov. Today: Technol.* **3**, 405–411.
232. Howe, T. J., Mahieu, G., Marichal, P., Tabruyn, T., and Vugts, P. (2006) Data reduction and representation in drug discovery. *Drug Discov. Today* **12**, 45–53.
233. Ivanenkov, Y. A., Savchuk, N. P., Ekins, S., and Balakin, K. V. (2009) Computational mapping tools for drug discovery. *Drug Discov. Today* **14**, 767–775.
234. Wagener, M., Sadowski, J., and Gasteiger, J. (1995) Autocorrelation of molecular surface properties for modeling corticosteroid binding globulin and cytosolic AH receptor activity by neural networks. *J. Am. Chem. Soc.* **117**, 7769–7775.
235. Bauknecht, H., Zell, A., Bayer, H., Levi, P., Wagener, M., Sadowski, J., and Gasteiger, J. (1996) Locating biologically

active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci.* **36**, 1205–1213.
236. Shneiderman, B. (1992) Tree visualization with tree-maps: 2-D space-filling approach. *ACM Trans. Graph.* **11**, 92–99.
237. Kibbey, C., and Calvet, A. (2005) Molecular property eXplorer: a novel approach to visualizing SAR using tree-maps and heatmaps. *J. Chem. Inf. Model.* **45**, 523–532.
238. Yamashita, F., Itoh, T., Hara, H., and Hashida, M. (2006) Visualization of large-scale aqueous solubility data using a novel hierarchical data visualization technique. *J. Chem. Inf. Model.* **46**, 1054–1059.
239. Lamping, J., Rao, R., and Pirolli, P. (1995) A focus + context technique based on hyperbolic geometry for visualizing large hierarchies, in *Proceedings of the SIGCHI conference on human factors in computing systems*, Denver, Colorado, United States, ACM Press/Addison-Wesley Publishing Co.
240. Agrafiotis, D. K., Bandyopadhyay, D., and Farnum, M. (2006) Radial clustergrams: visualizing the aggregate properties of hierarchical clusters. *J. Chem. Inf. Model.* **47**, 69–75.
241. Agrafiotis, D. K., and Xu, H. (2002) A self-organizing principle for learning nonlinear manifolds. *Proc. Natl. Acad. Sci. USA.* **99**, 15869–15872.
242. Agrafiotis, D. K., and Xu, H. (2003) A geodesic framework for analyzing molecular similarities. *J. Chem. Inf. Comput. Sci.* **43**, 475–484.
243. Agrafiotis, D. K. (2003) Stochastic proximity embedding. *J. Comput. Chem.* **24**, 1215–1221.
244. Patel, A., Chin, D. N., Singh, J., and Denny, R. A. (2006) Methods for describing a group of chemical structures. WO 2006023574.
245. Medina-Franco, J. L., Petit, J., and Maggiora, G. M. (2006) Hierarchical strategy for identifying active chemotype classes in compound databases. *Chem. Biol. Drug Des.* **67**, 395–408.
246. Schuffenhauer, A., Ertl, P., Roggo, S., Wetzel, S., Koch, M. A., and Waldmann, H. (2006) The scaffold tree – visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* **47**, 47–58.
247. Bemis, G. W., and Murcko, M. A. (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893.
248. Lipkus, A. H., Yuan, Q., Lucas, K. A., Funk, S. A., Bartelt, W. F., Schenck, R. J., and Trippe, A. J. (2008) Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry. *J. Org. Chem.* **73**, 4443–4451.
249. Schuffenhauer, A., Brown, N., Ertl, P., Jenkins, J. L., Selzer, P., and Hamon, J. (2007) Clustering and rule-based classifications of chemical structures evaluated in the biological activity space. *J. Chem. Inf. Model.* **47**, 325–336.
250. Agrafiotis, D. K., Shemanarev, M., Connolly, P. J., Farnum, M., and Lobanov, V. S. (2007) SAR maps: a new SAR visualization technique for medicinal chemists. *J. Med. Chem.* **50**, 5926–5937.
251. Kolpak, J., Connolly, P. J., Lobanov, V. S., and Agrafiotis, D. K. (2009) Enhanced SAR Maps: Expanding the Data Rendering Capabilities of a Popular Medicinal Chemistry Tool. J. Chem. Inf. Model. **49**, 2221–2230.
252. Smellie, A. (2007) General purpose interactive physico-chemical property exploration. *J. Chem. Inf. Model.* **47**, 1182–1187.
253. Krallinger, M., Erhardt, R. A.-A., and Valencia, A. (2005) Text-mining approaches in molecular biology and biomedicine. *Drug Discov. Today* **10**, 439–445.
254. Ananiadou, S., Kell, D. B., and Tsujii, J.-I. (2006) Text mining and its potential applications in systems biology. *Trends Biotechnol.* **24**, 571–579.
255. Erhardt, R. A. A., Schneider, R., and Blaschke, C. (2006) Status of text mining techniques applied to biomedical text. *Drug Discov. Today* **11**, 315–325.
256. Banville, D. L. (2006) Mining chemical structural information from the drug literature. *Drug Discov. Today* **11**, 35–42.
257. Banville, D. L. (2009) Mining chemical and biological information from the drug literature. *Curr. Opin. Drug Discov. Dev.* **12**, 376–387.
258. Banville, D. L., (Ed.) (2009) *Chemical Information Mining: Facilitating Literature-Based Discovery*, CRC Press, Boca Raton, FL.
259. RSC Prospect. http://www.rsc.org/Publishing/Journals/ProjectProspect/index.asp (accessed October 2, 2009).
260. Batchelor, C. R., and Corbett, P. T. (2007) Semantic enrichment of journal articles using chemical named entity recognition, in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions* Prague, Czech Republic, Association for Computational Linguistics, http://www.aclweb.org/anthology/W/W07/W07-1008.pdf (accessed October 2, 2009)

261. Corbett, P., Batchelor, C., and Teufel, S. (2007) Annotation of chemical named entities, in *BioNLP 2007: Biological, translational, and clinical language processing*, Prague, Czech Republic, Association for Computational Linguistics, http://www.aclweb.org/anthology/W/W07/W07-1008.pdf (accessed October 2, 2009).
262. Kidd, R. Prospecting for chemistry in publishing. paper given at ICIC 2008, Nice France October 2008. http://www.infonortics.eu/chemical/ch08/slides/kidd.pdf (accessed October 2, 2008).
263. RSC Ontologies. http://www.rsc.org/ontologies/ (accessed October 2, 2009).
264. Valko, A. T., and Johnson, A. P. (2009) CLiDE Pro: the latest generation of CLiDE, a tool for optical chemical structure recognition. *J. Chem. Inf. Model.* **49**, 780–787.
265. Eigner-Pitto, V., Eiblmaier, J., U. Frieske, U., Isenko, L., Kraut, H., Saller, H., and Loew, P. Mining for chemistry in text and images. A real world example revealing the challenge, scope, limitation and usability of the current technology. paper given at Fraunhofer-Symposium on Text Mining, Bonn, September 29–30, 2008. http://www.scai.fraunhofer.de/fileadmin/download/vortraege/tms_08/Valentina_Eigner_Pitto.pdf (accessed October 2, 2009).
266. Filippov, I. V., and Nicklaus, M. C. (2009) Optical structure recognition software to recover chemical information: OSRA, an open source solution. *J. Chem. Inf. Model.* **49**, 740–743.
267. Park, J., Rosania, G., Shedden, K., Nguyen, M., Lyu, N., and Saitou, K. (2009) Automated extraction of chemical structure information from digital raster images. *Chem. Cent. J.* **3**, 4.
268. Law, J., Zsoldos, Z., Simon, A., Reid, D., Liu, Y., Khew, S. Y., Johnson, A. P., Major, S., Wade, R. A., and Ando, H. Y. (2009) Route Designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *J. Chem. Inf. Model.* **49**, 593–602.
269. InfoChem. http://www.infochem.de (accessed October 2, 2009).
270. Matthews, E. J. Current Approaches for Toxicity Prediction, in *New Horizons in Toxicity Prediction. Lhasa Limited Symposium Event in Collaboration with the University of Cambridge.* http://www.qsarworld.com/lhasa_report1.php (accessed October 7, 2009).
271. Clark, R. D. (2009) At what point does docking morph into 3-D QSAR?, in *Abstracts of Papers, 237th ACS National Meeting, Salt Lake City, UT, United States, March 22–26, 2009.*
272. Scheiber, J., Chen, B., Milik, M., Sukuru, S. C. K., Bender, A., Mikhailov, D., Whitebread, S., Hamon, J., Azzaoui, K., Urban, L., Glick, M., Davies, J. W., and Jenkins, J. L. (2009) Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis. *J. Chem. Inf. Model.* **49**, 308–317.
273. Schreiber, S. L., Kapoor, T. M., and Wess, G., (Eds.) (2007) *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*, Wiley-VCH, Weinheim, Germany.