Review in Advance first posted online on February 26, 2015. (Changes may still occur before final publication online and in print.)

What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery

Edward O. Pyzer-Knapp, Changwon Suh, Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, and Alán Aspuru-Guzik

Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02143; email: aspuru@chemistry.harvard.edu

Annu. Rev. Mater. Res. 2015. 45:2.1–2.22

The *Annual Review of Materials Research* is online at matsci.annualreviews.org

This article's doi: 10.1146/annurev-matsci-070214-020823

Keywords

computational materials design, big data

Abstract

A philosophy for defining what constitutes a virtual high-throughput screen is discussed, and the choices that influence decisions at each stage of the computational funnel are investigated, including an in-depth discussion of the generation of molecular libraries. Additionally, we provide advice on the storing, analysis, and visualization of data on the basis of extensive experience in our research group.

1. INTRODUCTION

As a society, we categorize our history either by the prevalent materials (e.g., Bronze Age, Iron Age) or by the groundbreaking processes (e.g., the Industrial Revolution) related to their manufacture. It would not be unreasonable to say that we are now in the age of materials science—an age best categorized by the cornucopia of available materials made possible by a scientific method for discovery. The path to the present day, however, has not been simple or well defined. Many of the most significant materials discoveries have not been made by rational design but instead are a product of happenstance; the stars aligned, and the right person was in the right place at the right time. Perkin's mauve, vulcanized rubber, and Teflon are famous examples. This pattern should not be surprising, because the size of chemical space—recently estimated at $>10^{60}$ molecules (1)—makes any kind of rational global search challenging in the extreme.

Fortunately, global exploration is rarely required because we often have a good idea of the local area of chemical space in which we would like to explore. This reduces the size of the exploration from the order of 10^{60} molecules to somewhere on the order of 10^6. Although this is still far too many molecules for even the most advanced experimental screening techniques to consider, the massively parallel nature of the materials screening problem, coupled with recent advances in computer architecture and distribution techniques, has borne the concept of a virtual high-throughput screen, in which large libraries of molecules are analyzed by using theoretical techniques and are reduced to a small set of promising leads for experimental chemists to follow up on. Indeed, some have postulated that because of these developments, we are on the cusp of a golden age of materials discovery (2). Although the computational needs of the different areas that currently employ high-throughput screening techniques differ (**Figure 1**), a key set of philosophies is common throughout. We discuss the philosophy that defines a high-throughput screen, focusing on key areas such as library size, hierarchical techniques, and analysis methods. These techniques focus solely on the discovery of a new material, and not on the commercialization process, which is typically much longer.

2. THE PHILOSOPHY BEHIND HIGH-THROUGHPUT VIRTUAL SCREENING

Although high-throughput virtual screening has existed for a while (particularly in the pharmaceutical sciences), its definition is somewhat intractably tied to the capability of the state-of-the-art hardware at the time of calculation. We therefore propose that this field is better defined by the philosophy of the calculations undertaken rather than by the number, or intensity, of calculations (see sidebar, "Four Philosophies of High-Throughput Virtual Screening").

FOUR PHILOSOPHIES OF HIGH-THROUGHPUT VIRTUAL SCREENING

1. Significant timescale: Time-critical techniques require high-throughput solutions.
2. Automated techniques: High-throughput approaches require automation.
3. Data-driven discovery: Trends in the data are often as important as the data themselves.
4. Computational funnels: A funnel-like approach, in which only promising molecules are exposed to expensive calculation methods, allows for efficient deployment of computation.

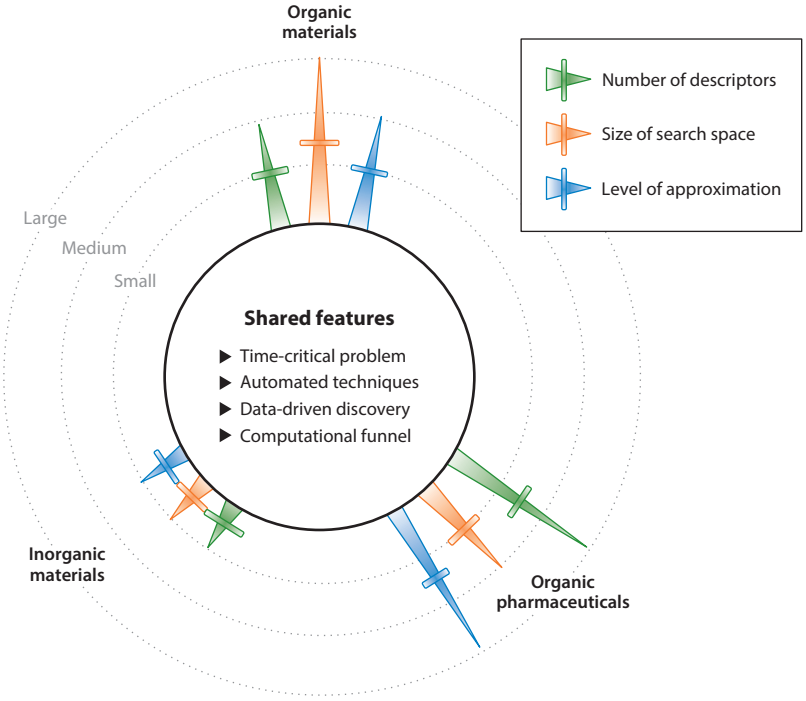

Figure 1

High-throughput screening techniques from many different areas share a core philosophy, despite having computational needs of different magnitudes.

2.1. Context

Although the context of the calculations and the chemical space differ between inorganic materials, organic materials, and organic pharmaceutical chemistry, high-throughput screens in all these areas share the same underlying philosophy. Together, these statements form a definition of a high-throughput screen. We expand upon each of these before reviewing their application in the literature as well as from our own research experiences.

2.2. Time-Critical Problems Require High-Throughput Solutions

Research moves at a variety of speeds. Whether due to pressure from competitors, financial motives, or social pressures, using a high-throughput technique to quickly focus on one particularly promising part of chemical space is often desirable. Additionally, some problems are intrinsically time critical. Our supplies of fossil fuels are finite and rapidly dwindling; however, our consumption continues to increase year after year. New technologies must be both developed and implemented soon to allow us to continue to feed our energy needs. Clearly, the timescale of studies such as these is of sufficient importance that it must influence the design of the investigation. High-throughput methodologies deliver greatest value in this area.

The timescale, whether required or desired, is one factor that we believe defines a high-throughput screen.

2.3. High-Throughput Approaches Require Automation

Manually performing a high-throughput screen is exceptionally challenging and often simply impossible. The generation, storing, and querying of the large volume of data require some degree of automation to be efficient, and each of these aspects is discussed in more detail in the upcoming sections. The increased use of combinatorial techniques adapted from life sciences for the generation of libraries has allowed for the creation of initial libraries of millions of candidate molecules, leaving a manual approach far beyond the ability of even the most fervent of investigators.

High-throughput screens require automation, especially in the early stages.

2.4. Computational Funnels Allow for Efficient Deployment of Computation

Frequently, the calculation of the property of interest in a screen is intrinsically too expensive for mass deployment over all potential molecules, thus potentially placing a direct calculation outside the realms of a high-throughput approach. One approach to avoiding this pitfall is to employ a computational funnel (**Figure 2**). Each level of the funnel represents a calculation with well-defined error bounds, and at each level structures are ruled out on the basis of selection criteria defined by the error bounds. Because each level is progressively more computationally intense, only the molecules most likely to be of interest are calculated with the most expensive methods, with each new level affording additional information about the molecule. The final "test" is an experimental fabrication of a device as close as possible to the expected running conditions—and because this process is both slow and expensive, the fewer candidate molecules that reach this stage, the better.

Following a computational funnel allows us to focus computational efforts on promising molecules.

3. MOLECULAR LIBRARIES

3.1. General Considerations

Every project that aims at a high-throughput screening of a section of molecular space must begin by selecting the candidate molecules to be investigated. This process is crucial because it puts boundaries to the possible outcomes from the start; the successful candidate can come only from the initial library or from its successive rounds of growth.

The generation of libraries is particularly important in exploratory research, when it is not known what type of molecule will solve a given problem, and somewhat less so when one is performing a thorough crib through some fixed, finite subsection of space. That is, generating novel lead backbones is a harder challenge than pursuing a complete set of substitution patterns through side-chain enumeration.

In molecular space, performing systematic explorations is particularly difficult: We lack any predefined magnitudes to survey chemical space so that we can move systematically along them and create a search grid or more sophisticated search pattern. Molecular structure obeys a large and complex set of rules that so far defies systematic exploration.

Despite the obvious, and subtle, dissimilarities with drug discovery in the pharma industry, high-throughput screening of organic materials has much to learn from its older and larger sibling. In the field of drug discovery, it is common to explore known chemical space for novel applications rather than try to create libraries de novo (3) or, at least, to perform variations on a known drug, after the maxim, "The most fruitful basis for the discovery of a new drug is to start with an old drug" (4). This policy is due to a combination of reasons, such as the large capital costs of developing new drugs, the high cost of late-stage failures, and the fact that structure–biological

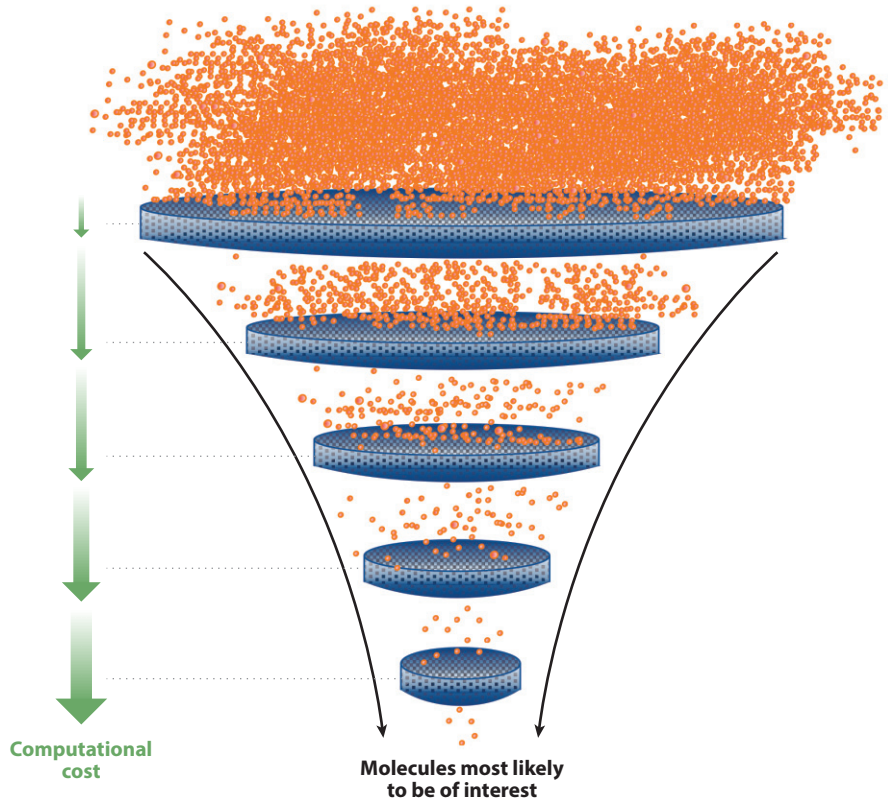

Figure 2

A computational funnel scheme. The increasingly strict filtering criterion eliminates many molecules that are not of interest and identifies the top-performing candidates in a virtual library.

activity relationships are usually much more obscure than structure–property relationships in the less mature area of organic materials screening.

Given the intrinsic impossibility of enumerating all the molecules in all but the smallest sections of molecular space, much effort is being put into methods to avoid explicit enumeration in high-throughput screening. Such approaches include optimization of potentials to make a rugged region of chemical space flatter (5), alchemical transformation on a given backbone (6), generation of aromatic rings that are electronically equivalent to a reference (7), stochastic generation of derivatives (8), recursive substructure searches (9), and morphing of starting molecules of interest (10).

However, more often than not, high-throughput-screening projects rely on explicitly enumerated libraries. Due to the need for diverse drug-like molecules, many examples of computer-generated molecular libraries have spawned in recent years, oriented mostly at compounds with potential biological activity. The chemical universe–generated databases aim at exhaustiveness: GDB-11, with 14 million molecules containing up to 11 atoms of C, N, O, and F (11); GDB-13, with under one billion molecules containing up to 13 atoms of C, N, O, S, and Cl (12); and GDB-17, with 166 billion molecules (13). Many other more oriented databases aim at exploring smaller subsets of larger molecules, such as tetrapyrrole macrocycles (14), natural product-like virtual libraries through recursive atom-based enumeration (15), or the recently reported ZINClick database of 16 million triazines (16).

3.2. Molecular Diversity

Molecular diversity is a vital concept in screening of molecular libraries that materials screening has also inherited from pharma. Much research in virtual molecular libraries has tried to maximize molecular diversity within a given library (17). The aim of diverse libraries is to do a homogeneous, multidimensional crib of molecular space, devoting little effort to exploring the neighborhood of areas already represented in the library.

The key issue, however, is that we lack absolute axes along which to survey molecular space (18), and thus we do not have universal metrics to assess similarity: A single isoelectronic alchemical substitution can completely disrupt the electronic structure of a molecule, whereas large variations in side chains can be relatively harmless to optical properties. Thus, what similarity metric to use, and ultimately how diverse a library is, is more an ad hoc decision based on chemical intuition than an intrinsic property of a library (19, 20).

3.3. Generation of Custom-Made Libraries

As mentioned above, the connections between molecular structure, electronic structure, and device properties in organic electronic materials are more straightforward than the structure–activity relationships in drug discovery, for which even knowing the macromolecular target hardly limits the choice of potential scaffolds. Because the CPU cost of computational methods in organic materials is quite high (see below), having as high a hit ratio as possible is important, and thus it is often more rewarding to generate custom libraries to trawl promising regions of molecular space than to use generic predefined ones. Many different codes have been reported for the assembly of virtual molecular libraries, oriented essentially toward lead discovery and optimization in pharma (21–30).

Library creation implies generating some or all of the possible combinations of a given set of fragments in a combinatorial way according to a set of rules. The fragments and the rules may borrow from experimentally feasible combinatorial synthesis schemes, or they may just be arbitrary schemes to explore chemical space with only an indirect connection to chemical synthesis.

In addition, it is necessary to set a limit when the growth of the molecule will come to an end. This termination procedure defines the maximum size of a given molecule and thus sets a ceiling for computational cost, helping to estimate the resources needed. In the most trivial case, a one-off combinatorial linkage, the maximum size is easily estimated with the maximum size of the two fragments to be joined. In more complex cases with variable growth steps, stopping points can be fixed at a given round of growth, at a maximum atom or electron count, or at a maximum molecular mass.

The challenge of selecting a slice of molecular space cannot be overstated. Deep chemical knowledge can be leveraged to generate libraries that explore novel regions and exploit promising ones while keeping the number of false positives as low as possible.

In addition, the fundamental target is to produce a library of molecules that fulfill some property requirement; that are also accessible, i.e., that are possible to synthesize; and that ultimately represent good value for investment—in both temporal and economic terms. These factors represent soft constraints that change not only between projects but also ultimately within a project: A more challenging synthesis can be pursued for a higher payoff (31). It is important to address these constraints at the earliest point: the construction of the molecular library. Substitution patterns in the fragments, and mode of growth, are key items here: A molecule is more likely to be synthesizable if the substitution pattern is the same in identical positions within a moiety.

The next challenge of molecular library generation is to encode these soft synthesis constraints into algorithmic molecular growth rules so that as little time as possible is wasted in synthetically

inaccessible regions of molecular space. Chemical intuition can be leveraged in the generation of libraries to maximize the chances of discovering molecules that are both synthetically accessible and useful for a given application. For instance, using constraints based on synthetic accessibility, Hutchison and colleagues (32, 33) reduced a set from potentially 800 million combinations to a mere 60,000 in a search for organic photovoltaic materials.

There is, however, a trade-off to encoding these soft constraints into hard algorithmic rules. Due to the fact that only what is in the library is screened, one risks leaving out of the process molecules that are harder to make but perhaps have game-changing properties: the high-risk, high-reward scenario. It would be desirable, then, to assess synthetic availability just like any other property along its own scale, and molecules can be judged globally. Despite large efforts to automatically assess synthetic availability (34–38), we have yet to reach the point at which these synthetic constraints can be extracted away from the library generation process and passed on to another score.

The computational efficiency can also be improved by hard coding some constraints into the library generation software. This is most commonly achieved by prohibiting and filtering out molecules carrying certain functional groups that may arise during the molecular generation procedure and that are either fundamentally unstable for the foreseen application or detrimental to the property of interest.

In the following paragraphs, we report three specific examples, currently being pursued by our own research group, of custom-generated libraries. These examples are in the areas of organic photovoltaics, small-molecule organic-based flow batteries, and organic light-emitting diodes.

3.4. Organic Photovoltaics

Given humanity's ever-increasing appetite for energy and the obvious drawbacks of conventional nonrenewable energy sources, developing materials to harvest solar energy has been one of the targets of high-throughput virtual screening. The Harvard Clean Energy Project (CEP) is an effort to discover the next generation of plastic solar cell materials (39). So far, more than 3 million molecules have been generated and a total of 300 million density functional theory (DFT) calculations have been performed to identify low-cost, high-efficiency organic photovoltaics.

All the molecules in the CEP library were grown from an initial selection of 26 fragments (40). The main source for this initial library is chemical intuition from experimental collaborators. This expertise was leveraged not only in the fragment selection, but also in the definition of the positions through which the fragments can be joined.

The strategy followed in this case included two possible fragment combinations. The first one, linking, created a single covalent bond between the two fragments undergoing the reaction. The second one, fusion, was used when both fragments included rings. This reaction made a final molecule in which a fused ring is created with the two original fragments sharing a covalent bond, as can be seen in **Figure 3**. Significantly, this second process requires the loss of two C atoms, so it is not related to a chemical process, as it is a purely computational construction. The cutoff for growth was the number of generations, with molecules all the way to tetramers (a combination of four original fragments) and a total number of molecules to be screened of more than 3 million.

3.5. Organic-Based Flow Batteries

Renewable energy sources such as sunlight or wind have a more unpredictable output than do traditional nonrenewable ones, and thus a shift toward greener energy sources will require

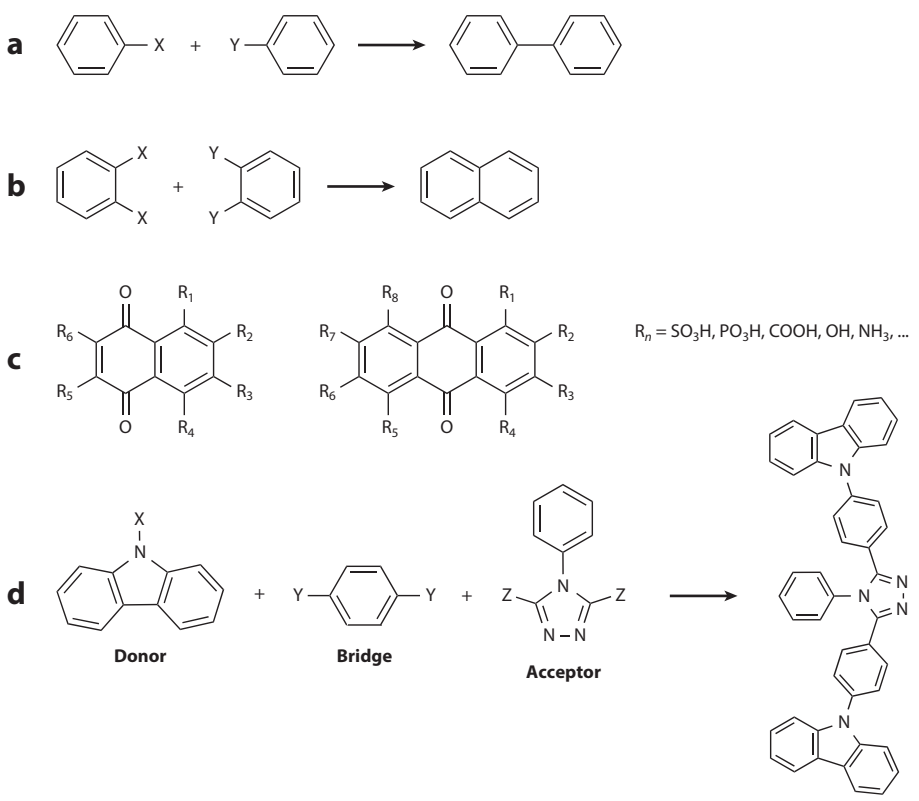

Figure 3

Reactions and combinations in virtual library enumeration. (*a*) The linking procedure used in the blue organic light-emitting diode (OLED) project. (*b*) The fusion procedure additionally utilized in the Clean Energy Project. (*c*) The enumeration of the different substitution positions considered in the organic-based flow battery project. (*d*) Combination of a donor molecule, a bridge molecule, and an acceptor molecule to give a potential blue OLED material.

developments in energy storage to compensate for the highs and lows in energy production. Flow batteries are a promising response, and Aziz and colleagues (41) have proposed the use of anthraquinone derivatives as electrolytes in flow batteries for massive storing of electrical energy. The use of organic redox species opens the doors to sustainable sourcing of the electrolytes. The choice of anthraquinone redox molecule was helped by a combinatorial screening involving R-group enumeration in multiple quinone backbones.

The goal of molecular generation in this project was to pursue a complete set of substitution patterns through R-group enumeration. The molecular frameworks were set from the start (benzo-, naptho-, and anthraquinone), and the full combinatorial space of substitution with one or multiple instances of each functional group was explored. These patterns were applied to several R groups, including sulfonate, hydroxy, and phosphonate. The nature of the side groups, mostly their electron withdrawing or electron donating and their polarity, tunes the two key chemical properties for these materials: redox potential and solubility.

The linking procedure described above is used to generate all possible combinations of core and R group. Libraries that explore full R-group enumeration can be very sizeable because they grow

factorially with the number of positions: In this case, all 1,2, 2,3, and 1,4 quinones are possible for a given ring backbone, and in each case, every remaining C atom in the quinone ring can bear a functional group (see **Figure 3**). A key issue to take into account is how many different R groups are combined in a given molecule. For a quinone with eight available C-H positions and a single R group, the number of possible molecules (excluding symmetry) is 8!; with the inclusion of a second R group, 8!·7!; and with inclusion of a third R group, 8!·7!·6!. The number becomes astronomical, even for a small number of different R groups. A very early cutoff has to be chosen regarding the maximum number of different groups in a given molecule. In this particular case, that number was limited to 2.

There is no need to select a termination strategy, because the maximum size is determined by the size of the core plus that of the substituents. In a first library (41), the substitution pattern explored was that of singly oxidized quinones either singly or fully substituted with one R group within a list of 14. This search was widened to benzo-, naphto-, and anthraquinones bearing two C = O groups and to any substitution pattern with any instances of a single R group (with a total of 3,037 unique substitution patterns).

3.6. Blue Organic Light-Emitting Material

Recent developments in thermally assisted, delayed fluorescence (TADF) have opened the door to novel classes of organic light-emitting diodes (OLED) (42, 43). These novel molecules exhibit low enough splitting between their lowest singlet state and triplet state that efficient thermal repopulation of the emissive singlet from the dark triplet is possible. Low-splitting excitations correspond to charge transfer states, and thus the basic TADF OLED must include electron donor and electron acceptor moieties, with some linker breaking the π conjugation.

TADF molecules have to satisfy a very specific donor–(bridge)–acceptor structure. As described above, it is advisable to encode as much chemical knowledge as possible into the library generation process to avoid spending valuable screening time in areas of chemical space that are barren by necessity. One can picture the inefficiency of a scenario in which fragments are combined without restrictions and only an analysis a posteriori would lead to the desired configuration.

Three successive strategies have been applied in the fragment selection. Initially, fragments that had been present in the OLED literature were selected. In a second effort, a new set of fragments not related to the OLED literature were selected and underwent a screen to trim undesirable candidates [in this case, according to the highest-occupied molecular orbital (HOMO) and lowest-occupied molecular orbital (LUMO) positions and the optical properties], and in a further step, to facilitate the synthesis of the final molecule, synthetic availability of each fragment was confirmed in the literature.

The third and final strategy included a random generation of fragments, creating one-, two-, and three-ring heterocycles with various amounts of N, O, and S and with a prescreen for electronic properties. This is a big-risk, big-reward scenario, in which completely new fragments not explored or synthesized before are studied.

As is the case with the organic flow battery project, only the linking described above was used. To fulfill a donor–acceptor strategy, the donor and acceptor space was first expanded with combinations of each of them with bridge fragments in a symmetrical fashion. This symmetry constraint, i.e., that analogous positions in a molecule grow in an identical way, is paramount for increasing the synthetic availability of molecules while restraining the combinatorial nature of the growth and reducing the computational cost of screening by orders of magnitude. In a final step, the donor and acceptor parts are combined with each other. For a small example, see **Figure 3**. A maximum molecular weight is determined to limit the size of the final molecules generated.

3.7. Other Considerations

Above, we very broadly explain three different strategies by which to achieve the generation of a set of molecules to be used for screening. There are a few ideas to be considered to further improve this process. We have seen the use of the concept of symmetry to guide synthesis. There are other ways to improve this process. We can forbid the creation of certain covalent bonds that may be very difficult or impossible to make. Similarly, if there are factors known about the physical processing of the final product (e.g., evaporation), it may be desirable to exclude molecules with certain limiting physical properties (e.g., molecular weight) from the molecular library.

The most crucial point, at least conceptually, is the fact that once the fragments and the way they combine are selected, we have limited ourselves to a very small area of the molecular space. No molecule outside that area will be screened. The decision made is a compromise between including a big part of the molecular space and keeping the computational expenses tractable. We should be able to establish a feedback loop to rethink the generated set with new fragments or combinatorial strategies when new information about the chemical space is available.

4. ON THE THEORETICAL CALCULATION OF MATERIALS PROPERTIES

The selection of simulation tools for high-throughput materials screening must be guided by clearly defined objectives in terms of the physical and chemical properties that are desired or necessary for the target technology. In other words, what are the optimization parameters, and what are the constraints? To answer these questions, it is often necessary to ask further questions about the desired physical properties of the material. For instance, under what conditions will the material be required to remain stable, will processing conditions place limits on molecular weight or on solubility, and what properties will be associated with a cheap and safe material? Is the target property thermodynamically or kinetically limited? Partnerships with industry may be useful in identifying the right constraints because synthesis or processing considerations may differ significantly between academic and industrial laboratory settings. The answers to these questions may guide the selection of the list of properties to screen for. For instance, a screen for battery electrolyte solvents might focus on a desired electrochemical stability window, melting and boiling point temperatures, viscosity, dielectric constant, Li-ion conductivity, and electronic conductivity ranges. Of these properties, only those that can be estimated with suitably cheap computational models can be chosen as initial screening criteria. In the realm of organic materials, this situation typically means that only single-molecule properties may be calculated, and solvent or solid-state effects must be estimated through an empirical model. In the example of electrolytes, the electrochemical stability window is an easy target property for quantum chemical simulations. Studies on a small family of related compounds will often guide the choices of what methodology is appropriate at different levels of a staged screen. We cannot emphasize enough that domain-specific knowledge must guide the selection of both search space and computational tools because judicious choices will vastly improve both the efficiency of the expended computer time and the viability of top-performing candidates.

Computational methods for screening materials properties might broadly be broken into the following simulation categories: quantum mechanical methods (semiempirical theory, DFT, or wave function theory), classical force field–based methods, and data-driven paradigms [encompassing quantitative structure–property relationships (QSPR), genetic algorithms, and machine-learning approaches]. By virtue of targeting high-throughput computation, computational cost sharply limits the available computational techniques. However, in screening we are interested

only in the ranking of candidates, and therefore systematic shifts between computational and experimental results need not be concerning as long as the faster method does not introduce so much error that we are unable to correctly identify the top hundred or so candidates (44). We demonstrate how the choice of property, and the wider context of the screen, can influence the methods used by looking in detail at three areas:

1. inorganic Li-ion batteries (e.g., the Materials Project),
2. organic photovoltaics (e.g., the CEP), and
3. organic-based flow batteries.

4.1. The Materials Project

The Materials Project is a high-throughput effort focused upon traversal of the inorganic materials space in search of novel battery materials. The project is run by Professor Gerbrand Ceder (MIT) and Dr. Kristin Persson (Lawrence Berkeley National Laboratory). To date, they have computed relevant properties of more than 80,000 materials and screened 25,000 of these materials for use as potential Li-ion batteries. This effort has so far used more than 15 million CPU hours of computational time at the National Energy Research Scientific Computing Center (NERSC).

The Materials Project mainly utilizes traditional computer sources (large supercomputer clusters). However, the results are distributed through a web portal (**http://www.materialsproject.org/**) and can be searched and analyzed by using a custom python module, PyMatGen (45).

Because the Materials Project is interested in electronic properties, a method based upon quantum mechanical principles is required. The high-throughput DFT computations were performed using the Vienna software package (VASP) (46), the projector augmented wave (PAW) pseudopotentials (47), and the generalized gradient approximation (GGA) (48). GGA was chosen because it represents a good compromise between speed and accuracy. To compensate for the known errors in the model due to electron self-interaction energy, calculations were performed within the DFT + U framework (49). Ceder et al. (50) note that the use of a hybrid functional—which reduces the effects of the self-interaction energy term—may increase the accuracy of the calculation and remove the need for operating within the DFT + U framework. These calculations were, however, considered too expensive for use in a high-throughput screen (50).

One particular method of calculation will sometimes have systematic failings for particular classes of molecules. The Materials Project aims to avoid any such biases within its data set by classifying materials into different classes, each with its own specific battery of associated calculations. The results from these calculations are then unified over the global population by using a set of reference reactions to connect results from different methods (51).

4.2. The Harvard Clean Energy Project

The CEP is a search for molecules with the potential to be employed in organic photovoltaic devices (39, 52). It is unique among materials high-throughput screening projects in the scale at which it utilizes distributed computing through the World Community Grid (**https://secure.worldcommunitygrid.org/index.jsp**). The computing power of the project is estimated to be on the order of a fully utilized 6,000–7,000-core cluster, and the amount of harvested CPU time can be seen in **Figure 4**.

With such a large amount of computing power available, an approach slightly different from that of the Materials Project was taken. Instead of one fast functional (GGA), the CEP calculates properties of molecules with a range of different functionals both to reduce systematic errors that

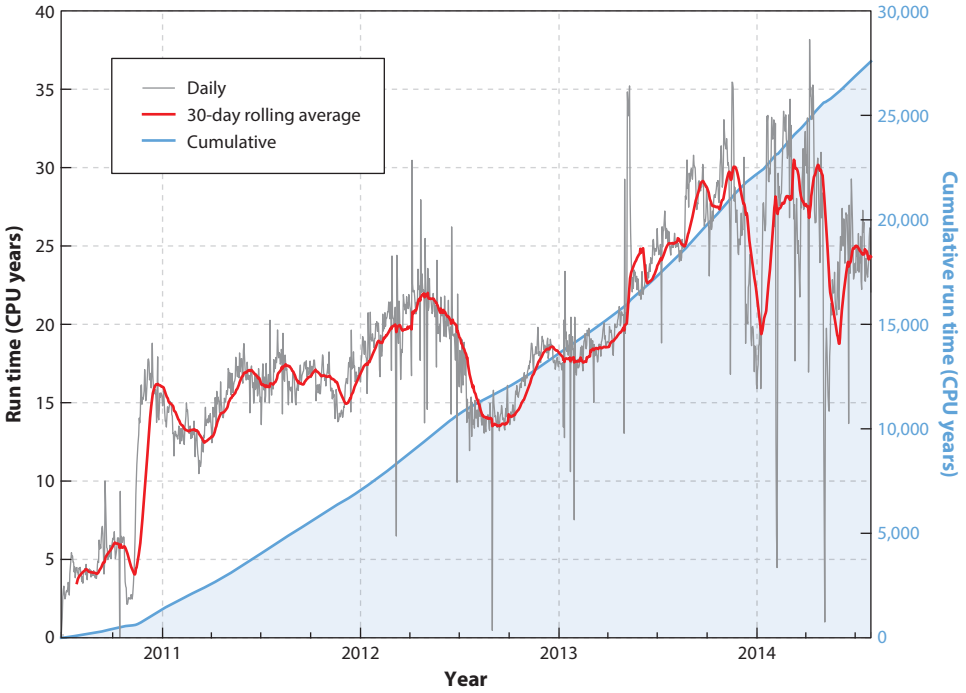

Figure 4

The amount of CPU harvested over the course of the Clean Energy Project calculated as 1-CPU run-time equivalent.

occur in particular functionals and to investigate how the choice of functional affects the values computed. The basic workflow consists of two types of jobs: an optimization of the molecular geometry and a single-point calculation on that optimized geometry. Even with the vast resources available, optimizations using all the different functionals were not considered to be a good deployment of computational power, and so the geometry as optimized using the BP86 functional was used for all single-point calculations.

For solar cell performance, the electronic structure is not the property of interest; rather the way in which the electronic structure interacts with photons of sunlight provides the true ranking metric. For this, the Scharber model of the power conversion efficiency (53) was used to rank molecules. The Scharber model is a specialized version of the Shockley–Queisser model (54) for OPVs and is based upon the energies of the HOMO and the LUMO—both of which are easily calculable properties.

Due to its large size (more than 3 million molecules and more than 300 million quantum chemical calculations), the CEP Database (**https://cepdb.molecularspace.org/**) provides an ideal test bed for data-driven approaches such as cheminformatics and machine learning. These methods represent the potential to quickly and rigorously develop, from existing data, QSPR-type models, which can then be used to focus calculation efforts upon promising areas of chemical space. Olivares-Amaya et al. (40) have used cheminformatics descriptors to this effect, calculating the current–voltage properties of more than 2.5 million molecules by using linear regression descriptor models.

4.3. Organic-Based Flow Batteries

The search for a metal-free flow battery by Aziz and colleagues (41) represents a good example of how a high-throughput virtual screen can complement an experimental study, improving the efficiency of optimizing the molecule within a chemical space.

An essential component of this new generation of aqueous redox flow battery is a unique quinone molecule, 9,10-anthraquinone-2,7-disulfonic acid (AQDS). The small, electroactive, and water-soluble AQDS molecule that is used at the negative side of a flow battery was identified by using quantum chemical calculations on a pool of approximately 10,000 candidate molecules. Computational studies focused on investigating how changing the substitution patterns of AQDS would affect two key properties:

1. the redox potential, E^0, of the quinone–hydroquinone couples and
2. the solvation free energy, G^0_{solv}, of quinones in water.

For a high-throughput search such as this one, the generation of three-dimensional structures (conformers) are an important component of calculation workflow. As with any calculation, there exists a trade-off between the computational cost of a calculation and accuracy. Because the generation of many conformers represents a more complete exploration of the energy landscape of a molecule, an additional criterion is introduced: the completeness of the search.

With a high-throughput screening, the number of molecules to be processed dictates the amount of time that the algorithm can spend on each molecule. Conformer generation algorithms can generally be split into two categories:

1. physically motivated generators [e.g., a low-mode generator (55)] and
2. rule-based generators [e.g., CORINA (56)].

The most reliable way to have an acceptably complete search (if one ignores the trivial case of simply varying all torsion angles and calculating the energy, which is simply too costly of a method to use in anything other than the simplest of cases) is to use a physically motivated generator, such as the low-mode conformer generator implemented in MacroModel (Schrödinger). These techniques have shown good results in many different circumstances but are slow to run. In general, the rule-based conformer generators are much faster than their physically based counterparts. The price that comes with that speed increase is that, due to their very nature, the rule-based conformer generators are applicable only to molecules that are similar to those that were used to develop the rules on which the generators are based. Because the molecules within Aziz et al.'s (41) search were similar, these researchers determined that a rule-based search was acceptable, and starting geometries were improved by minimization using the DREIDING force field (57), from which a low-energy selection was then reminimized using DFT (GGA/PBE)—which, as discussed in the previous example, represents a good compromise between speed and accuracy. This project continued to use the concept of the computational funnel with its approach to calculating E^0 and G^0_{solv}—the properties that are of primary interest for the development of flow batteries utilizing water-soluble, electroactive molecules.

5. CONSIDERATIONS FOR HIGH-THROUGHPUT VIRTUAL SCREENING

5.1. Deployment of a High-Throughput Virtual Screen

Having decided that high-throughput virtual screening is the correct solution to a particular scientific problem, we now shift focus to the deployment of the calculations that will produce the data of interest. The university, or company, computer cluster is the traditional locale for

these computations. The computational capability across these types of machines is typically very isotropic and reliable, with a file system that allows a large number of intense calculations to be run at the same time. The one major downside to using these resources is that they are typically shared among many users, limiting the time that any one user can get for her own projects. One other alternative option that has gained traction recently is the use of distributed computing. This approach uses idle time on the computers of volunteers to achieve a facsimile of the computer cluster by borrowing computer time from these transient nodes. Popularized by the Folding at Home (**http://folding.stanford.edu**) and SETI at Home (**http://setiathome.berkeley.edu**) projects, this approach has yielded good results for a wide variety of projects, especially through IBM's World Community Grid initiative. As demonstrated in **Figure 4**, this approach can result in significant amounts of CPU resource—which is invaluable to a high-throughput virtual screen. Of course, there are downsides to this approach, which result mainly from the fact that calculations are being performed upon computers whose main purpose is not simply to compute, but so long as the project requirements are not too strenuous, this approach remains valid. Indeed, we are starting to see universities utilize it for their own research (see, e.g., **http://www.ucs.cam.ac.uk/scientific/camgrid**).

5.2. Dealing with Data from High-Throughput Screens

Data that are a product of a high-throughput screen produce a series of challenges, which may be unfamiliar to scientists. Although Section 6 is dedicated to the analysis and visualization of large data sets, it is also important to consider how these data are stored and accessed.

Databases offer an attractive solution to both the storage of and the structured querying of data. However, there are many flavors of database architectures, and choosing the right one for your needs is crucial.

In general, databases are split into two camps: relational databases based upon Structured Query Language (SQL) and nonrelational databases (NoSQL). Each offers distinct pros and cons that should be weighed when one is deciding how to proceed. SQL databases offer complete transactional integrity but lack flexibility. In contrast, NoSQL databases offer a much more flexible structure but do not guarantee transactional integrity. Additionally, the object-oriented nature of NoSQL databases makes them more intuitive for storing different types of data that are linked to one parent. **Figure 5** presents a basic schema that shows how the flexible NoSQL format allows for the storage of different types of data. Due to its flexible framework and data types, the same underlying model has been successfully applied to projects of varying sizes and varying needs, and so, within our group, we favor the use of a NoSQL database (MongoDB) for the storage of the vast majority of our information.

Figure 6 summarizes the main characteristics of SQL and NoSQL databases. A well-implemented NoSQL database will outperform a poorly implemented SQL database, and vice versa.

One major benefit of interacting with data in a database format, whatever the flavor, is that it gives the scientist access to a wide range of tools specifically designed for analyzing large numbers of data. All the database architectures have their own shell for querying and allow access through a variety of popular scripting languages; large-scale data analysis has never been easier.

Another advantage of NoSQL databases is the fact that their performance scales both horizontally and vertically (**Figure 7**). Whereas SQL databases generally need to be stored on one machine, NoSQL databases were designed to be split over many machines—a process known as sharding. The process of sharding spreads data across shards (servers), and because NoSQL architectures natively implement this process, most support balancing and query loading, resulting in good database performance with minimal maintenance costs. Because performance can be gained

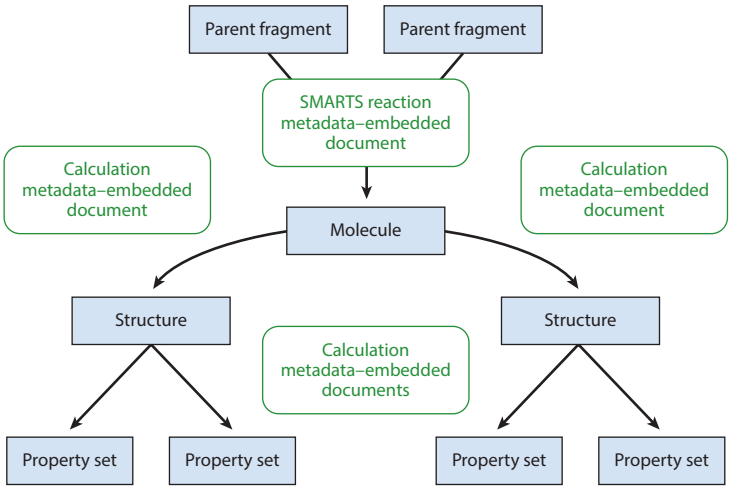

Figure 5

A basic schema showing how different types of data (molecular descriptors, geometries, properties, calculation metadata) can coexist easily within a NoSQL format. Arrows represent bidirectional links between documents and are embedded within the parent document and the child document. Additional metadata are embedded within documents, which improves performance by reducing the number of steps required in a query.

by simply adding more machines to the system, this option is cheaper than purchasing increasingly powerful machines as your database grows.

6. ON THE ANALYSIS OF LARGE NUMBERS OF CHEMICAL DATA

Analyzing chemical data generated from high-throughput screening in a high-throughput manner is highly desirable. The primary targets using high-throughput data analysis are for (*a*) the suggestion, prioritization, and identification of top candidates for targeted synthesis and (*b*) uncovering

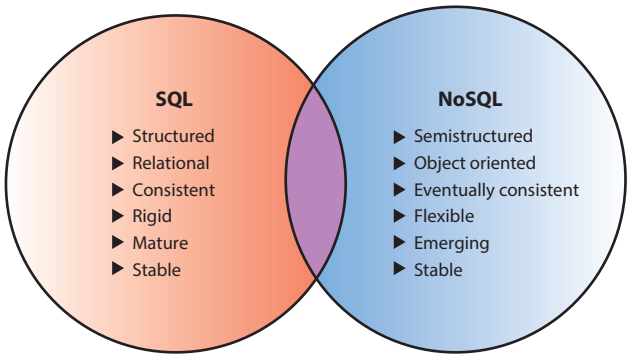

Figure 6

A comparison between SQL and NoSQL architectures. Although SQL databases allow for transactional integrity, we believe that the enhanced flexibility of NoSQL databases makes them the ideal choice for high-throughput virtual screening because they can be easily modified to adapt to changing data and requirements.

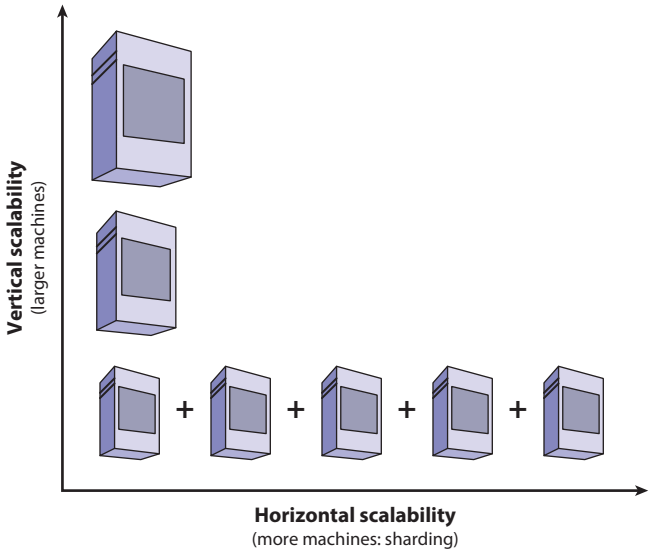

Figure 7

A diagrammatic representation of vertical and horizontal scaling in database architectures. Because there is a nonlinear scaling between computer power and cost, a vertical solution is more expensive than its horizontal counterpart.

sophisticated knowledge described as quantitative structure–activity relationships (QSAR)/QSPR toward rationalization of selected candidates (3, 58). More importantly, using high-throughput data analysis creates unprecedented insight for future experiments. Both the generation of large-scale data and fast, yet accurate, data analysis in an unbiased manner are equally important in high-throughput screening (59, 60).

Due to the sheer numbers of data produced in a high-throughput screen, using traditional data analysis techniques will impose a significant bottleneck upon the screening procedure. This will in turn impose restrictions on the ability of the screening cycle to adapt to emerging trends and results (61). We face exponentially growing requirements around organizing, summarizing, and interpreting such large numbers of high-throughput data with respect to the vast chemical space to explore. In particular, there are two challenges in high-throughput data analysis. First, the data we collect are increasingly voluminous, high dimensional, complex, noisy, and diverse. Second, the chemical space to explore is essentially infinite, and even a well-designed chemical library cannot cover all possible chemistries (62).

Currently, chemical data being generated through high-throughput screening overwhelm traditional data analysis, undermining our ability to perform interactive and exploratory analysis and visualization of these data in postprocessing. This situation will be exacerbated in the near future because the number of data produced will grow by orders of magnitude due to the ever-increasing hardware capability of computational resources. Although high-throughput screening still needs to provide the role of traditional data analysis, which scrutinizes each individual data point, a further complication is the removal of bottlenecks to suggesting candidates and extracting useful information such as QSAR/QSPR by treating the data as a whole. To solve the issues addressed here, there are four main components to consider and perform in high-throughput-screening data analysis: processing, mapping and visualization, interpretation, and modeling.

First, with regard to data processing, high-throughput data analysis starts with ensuring the quality of data that one collects for further discovery procedures. Multiple tasks such as cleaning,

organization, normalization, and outlier detection need to be performed before a full data exploration. When one is summarizing data, statistics, including analyses of systematic error, correlation, and associations, must be combined with simple visualizations such as heat maps (63). Before performing additional data analysis, one should consider molecular descriptors that are encoded representations of molecules and useful for construction of QSAR/QSPR models (60, 64). Unlike high-throughput data analysis for inorganic materials, analysis of chemical data has distinct features in the development of descriptors. With the aid of graph theory, for instance, fingerprints and fragments have become the preferred descriptors for effective exploration of chemical space (3, 52).

Second, the approach of data mapping and visualization is one of the crucial components in high-throughput data analysis because it is a direct way to depict chemical data and/or mined results with respect to chemical space to get insight into QSAR/QSPR (65, 66). However, simultaneous mapping of information is not always simple due to the nature of multidimensionality stemming from enumerated molecules, rules for bonds, functional groups, their chemical properties, and so forth. According to the similar-property principle (67, 68), similar chemical structures should have similar physicochemical properties. In that sense, it is crucial to choose a proper measure of distance in a high-dimensional space to logically place molecules in property spaces defined by proper coordinates in high-dimensional visualization for visual mining (66, 69–71). There are two broad categories of visualization for chemical information: direct visualization without data treatment and direct visualization with data-mined results. In the first category are plot matrix, parallel coordinates, heat maps, and other approaches (65). The typical approach for the latter category includes dimensionality reduction to decrease high dimensions to manageable levels. Such reduction can be achieved by linear and nonlinear dimensionality reduction by using, for example, principal-component analysis and diffusion map embedding (72).

Third, with regard to data interpretation, given the large quantity of data generated within a vast chemical space during high-throughput screening, it is a natural choice to use data mining for identifying hot spots (e.g., interesting regions) where we are likely to find more candidates with desired functionalities from a vast chemical space (3, 60). Data mining provides a flexible computational path for meeting the need to explore large amounts of chemical information. We often use various methods such as classification, clustering, and prediction.

Finally, with regard to data modeling for calibration, as mentioned above, despite our best methods, experimental values are often not reproduced by computational techniques due to the inclusion of some systematic error. One way to correct for this is to calibrate the calculated results to experimental values. This approach results in data that are more intuitively analyzed by experimental collaborators.

The challenge in conducting high-throughput screening is ensuring the pace of exploration of search space while keeping high levels of accuracy of calibration models (73). Given the two critical factors of speed and accuracy, high-throughput screening mainly utilizes two types of approaches in advanced calibration modeling: hard modeling and soft modeling. To understand materials behaviors in a unified way, hard modeling captures different length scales of materials behaviors with chemistry- and physics-based theories and integrates such information (74). Although it provides highly accurate results, such modeling generally incurs a high computational cost. Exemplary approaches include ab initio calculations and thermodynamic modeling.

In the organic chemistry community, soft modeling has been a powerful approach not only for enhancing the accuracy of hard modeling (75, 76) but also for making fast and accurate property predictions (77). Soft modeling is based on statistical learning methods to seek heuristic relationships between data (74) and often uses developed descriptors as well as knowledge extracted through the previous modeling tasks to construct a cheap yet robust model enabling the establishment and deeper understanding of QSAR/QSPR (58). It includes regressions, artificial

neural networks, multivariate analysis, and other machine-learning algorithms (78). Unlike hard modeling, predictive soft modeling is particularly valuable when physical or chemical models are not available. Models from soft modeling are relatively cheaper to construct than those fully generated from expensive quantum chemical calculations, allowing for accelerated screening procedures by replacing a huge number of such theoretical investigations with heuristically developed QSAR/QSPR.

A good example of predictive calibration approach is the application of Gibbs energy relationships to correlate electrode potentials of quinone–hydroquinone couples (41). Inspired by Dewar & Trinajstic's (79) early work, Huskinson et al. (41) computed the differences in gas-phase energy between oxidized and reduced states of quinone couples and successfully correlated those with measured redox potential values.

Robust calibration models should be developed from unskewed subspaces and should be generalized to appropriately cover other possible chemistries. In other words, to ensure a higher prediction power for a wide range of interesting candidates, the calibrated data need to be distributed in a proper range to cover a larger chemical space of molecules to explore. Moreover, the quality of the training set determines the accuracy of the calibration model when one screens a large number of new hypothetical molecules with limited or no experimental data in particular.

High-throughput analysis of chemical data is a critical field of study in chemistry. It has emerged to address the issues associated with larger, more complex data sets. With increased understanding of relationships between structure and property through high-throughput data analysis, the most concrete outcome of data-driven models is to direct future experiments to discover high-performance materials. More abstractly, application of such data-driven models will greatly enhance our understanding of basic physical and chemical principles (80). To those ends, data fusion and informatics platforms are important pieces that take greater advantage of such knowledge. Data fusion is an indispensable procedure in cutting-edge high-throughput data analysis to link structures and properties (81). It can be more effectively completed when high-throughput data analysis within the proper informatics platforms is performed (61).

7. FUTURE DIRECTIONS

With the constant onslaught of new and improved computational hardware, and the increased uptake of deployment techniques such as distributed computing, we strongly believe that high-throughput virtual screening has an important role to play in the future of materials science. A key factor in this regard will be the resurgence of cheap, approximate techniques such as semiempirical quantum methods and QSPR approaches. Although the lower accuracy of these methods has resulted in a downturn in their usage in isolation, they gain a new value when included in a computational funnel, because only relative values are important and known errors can be tolerated.

We predict that there will also be a sharp increase in the use of machine-learning techniques to act as a fast approximation for materials properties. The ability to use these techniques to exploit complex relationships between seemingly unconnected descriptors, and the fact that these models can be easily trained against a small subset of chemical space that is directly relevant to the specific problem at hand, makes these techniques ideal for attacking these problems. Additionally, Bayesian methods produce not only knowledge of the result of the model, but also confidence in that answer. This property can in turn be exploited to train the model on the fly, significantly increasing its value.

For these models to work, there must be good-quality experimental results to calibrate against. Thus, experiments must be performed in both an exploitative manner and an exploratory manner. That is to say, experiments that increase the knowledge of the local chemical space are as important

as those that are focused on optimizing properties. This approach can exist only with continued implementation of the ideas of the Materials Genome Initiative, which aims to enable experimental and theoretical teams to work together to reap the benefits from the increased efficiency derived from use of a high-throughput virtual screen.

With regard to experimental data, automated methods to collect the data and to classify them with respect to experimental condition and measurement type are required for further progress. Automated statistical methods to aid the assessment of reported versus actual experimental error bars will allow for better calibration of theory to experiment.

Finally, the development of software tools aimed at collaboration between theoreticians and experimentalists who are poring over large data sets of high-throughput virtual screens, and iteration of synthesis and computation, will help bring screening methods to many more communities beyond the current group of scientists that use them.

SUMMARY POINTS

1. A high-throughput virtual screen is best defined by the philosophy employed in approaching the problem.
2. Library generation is a compromise between including a big part of the molecular space and keeping the computational expenses tractable; a feedback loop between theory and experiment is recommended to keep the search in productive areas of chemical space.
3. Both the cost and the accuracy of calculations in a high-throughput virtual screen should be considered. Cheap methods have significant value in minimizing the number of molecules that are calculated by using high-level methods.
4. Identifying trends in the data is as important as identifying specific results.
5. Calibrating results to experimental data can overcome deficiencies in specific methods.
6. Exploratory, as well as exploitative, experimental results are crucial for the long-term success of high-throughput virtual screening.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank Martin Blood-Forsythe and Suleyman Er for helpful discussions. A.A-G. also acknowledges the following funding sources: Samsung Electronics Co., Ltd.; the Department of Energy through grant DE-SC0008733; and ARPA-E (Advanced Research Projects Agency–Energy) through grant DE-AR0000348.

LITERATURE CITED

1. Reymond J-L, van Deursen R, Blum LC, Ruddigkeit L. 2010. Chemical space as a source for new drugs. *Med. Chem. Commun.* 1:30
2. Cedar G, Persson K. 2013. How supercomputers will yield a golden age of materials science. *Sci. Am.*, Nov. 19

3. Lipinski C, Hopkins A. 2004. Navigating chemical space for biology and medicine. *Nature* 432:855–61
4. Wermuth C. 2006. Selective optimization of side activities: the SOSA approach. *Drug Discov. Today* 11:160–64
5. Wang M, Hu X, Beratan DN, Yang W. 2006. Designing molecules by optimizing potentials. *J. Am. Chem. Soc.* 128:3228–32
6. Balawender R, Welearegay MA, Lesiuk M, Proft FD, Geerlings P. 2013. Exploring chemical space with the alchemical derivatives. *J. Chem. Theory Comput.* 9:5327–40
7. Tu M, Rai BK, Mathiowetz AM, Didiuk M, Pfefferkorn JA, et al. 2012. Exploring aromatic chemical space with NEAT: Novel and Electronically equivalent Aromatic Template. *J. Chem. Inform. Model.* 52:1114–23
8. Virshup AM, Contreras-García J, Wipf P, Yang W, Beratan DN. 2013. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.* 135:7296–303
9. Ehrlich HC, Henzler AM, Rarey M. 2013. Searching for recursively defined generic chemical patterns in nonenumerated fragment spaces. *J. Chem. Inform. Model.* 53:1676–88
10. Hoksza D, Škoda P, Voršilák M, Svozil D. 2014. Molpher: a software framework for systematic chemical space exploration. *J. Cheminform.* 6:7
11. Fink T, Bruggesser H, Reymond J-L. 2005. Virtual exploration of the small-molecule chemical universe below 160 Daltons. *Angew. Chem. Int. Ed.* 44:1504–8
12. Blum LC, Reymond J-L. 2009. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* 131:8732–33
13. Ruddigkeit L, van Deursen R, Blum LC, Reymond J-L. 2012. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inform. Model.* 52:2864–75
14. Taniguchi M, Du H, Lindsey JS. 2011. Virtual libraries of tetrapyrrole macrocycles. combinatorics isomers, product distributions, and data mining. *J. Chem. Inform. Model.* 51:2233–47
15. Yu MJ. 2011. Natural product-like virtual libraries: recursive atom-based enumeration. *J. Chem. Inform. Model.* 51:541–57
16. Massarotti A, Brunco A, Sorba G, Tron GC. 2014. ZINClick: a database of 16 million novel patentable, and readily synthesizable 1,4-disubstituted triazoles. *J. Chem. Inform. Model.* 54:396–406
17. Koutsoukas A, Paricharak S, Galloway WRJD, Spring DR, IJzerman AP, et al. 2014. How diverse are diversity assessment methods? A comparative analysis and benchmarking of molecular descriptor space. *J. Chem. Inform. Model.* 54:230–42
18. Roth HJ. 2005. There is no such thing as 'diversity'! *Curr. Opin. Chem. Biol.* 9:293–95
19. Riniker S, Landrum GA. 2013. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminform.* 5:26
20. Maggiora G, Vogt M, Stumpfe D, Bajorath J. 2014. Molecular similarity in medicinal chemistry. *J. Med. Chem.* 57:3186–204
21. Gillet V, Johnson A, Mata P, Sike S, Williams P. 1993. SPROUT: a program for structure generation. *J. Comput. Aided Mol. Des.* 7:127–53
22. Pearlman D, Murcko M. 1996. CONCERTS: dynamic connection of fragments as an approach to de novo ligand design. *J. Med. Chem.* 39:1651–63
23. Schneider G, Lee M, Stahl M, Schneider P. 2000. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. *J. Comput. Aided Mol. Des.* 14:487–94
24. Gillet V, Willett P, Fleming P, Green D. 2002. Designing focused libraries using MoSELECT. *J. Mol. Graph. Model.* 20:491–98
25. Vinkers H, de JM, Daeyaert F, Heeres J, Koymans L, et al. 2003. SYNOPSIS: SYNthesize and OPtimize System in Silico. *J. Med. Chem.* 46:2765–73
26. Brown N, McKay B, Gasteiger J. 2004. The de novo design of median molecules within a property range of interest. *J. Comput. Aided Mol. Des.* 18:761–71
27. Nicolaou C, Brown N, Pattichis C. 2007. Molecular optimization using computational multi-objective methods. *Curr. Opin. Drug Discov. Devel.* 10:316–24
28. Liu Q, Masek B, Smith K, Smith J. 2007. Tagged fragment method for evolutionary structure-based de novo lead generation and optimization. *J. Med. Chem.* 50:5392–402

29. Dey F, Caflisch A. 2008. Fragment-based de novo ligand design by multiobjective evolutionary optimization. *J. Chem. Inform. Model.* 48:679–90
30. Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, et al. 2012. Automated design of ligands to polypharmacological profiles. *Nature* 492:215–20
31. Osedach TP, Andrew TL, Bulović V. 2013. Effect of synthetic accessibility on the commercial viability of organic photovoltaics. *Energy Environ. Sci.* 6:711–18
32. O'Boyle NM, Campbell CM, Hutchison GR. 2011. Computational design and selection of optimal organic photovoltaic materials. *J. Phys. Chem. C* 115:16200–10
33. Kanal IY, Owens SG, Bechtel JS, Hutchison GR. 2013. Efficient computational screening of organic polymer photovoltaics. *J. Phys. Chem. Lett.* 4:1613–23
34. Bertz SH. 1981. The first general index of molecular complexity. *J. Am. Chem. Soc.* 103:3599–601
35. Boda K, Johnson A. 2006. Molecular complexity analysis of de novo designed ligands. *J. Med. Chem.* 49:5869–79
36. Bonnet P. 2012. Is chemical synthetic accessibility computationally predictable for drug and lead-like molecules? A comparative assessment between medicinal and computational chemists. *Eur. J. Med. Chem.* 54:679–89
37. Podolyan Y, Walters MA, Karypis G. 2010. Assessing synthetic accessibility of chemical compounds using machine learning methods. *J. Chem. Inform. Model.* 50:979–91
38. Warr WA. 2014. A short review of chemical reaction database systems computer-aided synthesis design, reaction prediction and synthetic feasibility. *Mol. Inf.* 33:469–76
39. Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, Amador-Bedolla C, Sánchez-Carrera RS, et al. 2011. The Harvard Clean Energy Project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.* 2:2241–51
40. Olivares-Amaya R, Amador-Bedolla C, Hachmann J, Atahan-Evrenk S, Sánchez-Carrera RS, et al. 2011. Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy Environ. Sci.* 4:4849–61
41. Huskinson B, Marshak MP, Suh C, Er S, Gerhardt MR, et al. 2014. A metal-free organic–inorganic aqueous flow battery. *Nature* 505:195–98
42. Goushi K, Yoshida K, Sato K, Adachi C. 2012. Organic light-emitting diodes employing efficient reverse intersystem crossing for triplet-to-singlet state conversion. *Nat. Photonics* 6:253–58
43. Zhang Q, Li B, Huang S, Nomura H, Tanaka H, Adachi C. 2014. Efficient blue organic light-emitting diodes employing thermally activated delayed fluorescence. *Nat. Photonics* 8:326–32
44. Korth M. 2014. Large-scale virtual high-throughput screening for the identification of new battery electrolyte solvents: evaluation of electronic structure theory methods. *Phys. Chem. Chem. Phys.* 16:7919–26
45. Ong SP, Richards WD, Jain A, Hautier G, Kocher M, et al. 2013. Python Materials Genomics (pymatgen): a robust open-source python library for materials analysis. *Comput. Mater. Sci.* 68:314–19
46. Kresse G, Furthmüller J. 1996. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* 6:15–50
47. Blöchl PE. 1994. Projector augmented-wave method. *Phys. Rev. B* 50:17953–79
48. Perdew JP, Burke K, Ernzerhof M. 1996. Generalized gradient approximation made simple. *Phys. Rev. Lett.* 77:3865–68
49. Anisimov VI, Zaanen J, Andersen OK. 1991. Band theory and Mott insulators: Hubbard *U* instead of Stoner *I*. *Phys. Rev. B* 44:943–54
50. Jain A, Hautier G, Moore CJ, Ong SP, Fischer CC, et al. 2011. A high-throughput infrastructure for density functional theory calculations. *Comput. Mater. Sci.* 50:2295–310
51. Jain A, Ong SP, Hautier G, Chen W, Richards WD, et al. 2013. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* 1:011002
52. Hachmann J, Olivares-Amaya R, Jinich A, Appleton AL, Blood-Forsythe MA, et al. 2014. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry—the Harvard Clean Energy Project. *Energy Environ. Sci.* 7:698–704
53. Scharber MC, Mühlbacher D, Koppe M, Denk P, Waldauf C, et al. 2006. Design rules for donors in bulk-heterojunction solar cells—towards 10% energy-conversion efficiency. *Adv. Mater.* 18:789–94

54. Shockley W, Queisser HJ. 1961. Detailed balance limit of efficiency of *p-n* junction solar cells. *J. Appl. Phys.* 32:510
55. Kolossváry I, Guida WC. 1996. Low mode search. An efficient automated computational method for conformational analysis: application to cyclic and acyclic alkanes and cyclic peptides. *J. Am. Chem. Soc.* 118:5011–19
56. Sadowski J, Gasteiger J, Klebe G. 1994. Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J. Chem. Inform. Model.* 34:1000–8
57. Mayo SL, Olafson BD, Goddard WA. 1990. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* 94:8897–909
58. Parker CN, Shamu CE, Kraybill B, Austin CP, Bajorath J. 2006. Measure, mine, model, and manipulate: the future for HTS and chemoinformatics? *Drug Discov. Today* 11:863–65
59. Tamura SY, Bacha PA, Gruver HS, Nutt RF. 2002. Data analysis of high-throughput screening results: application of multidomain clustering to the NCI anti-HIV data set. *J. Med. Chem.* 45:3082–93
60. Harper G, Pickett SD. 2006. Methods for mining HTS data. *Drug Discov. Today* 11:694–99
61. Ling X. 2008. High throughput screening informatics. *Comb. Chem. High Throughput Screen.* 11:249–57
62. Medina-Franco J, Martínez-Mayorga K, Giulianotti M, Houghten R, Pinilla C. 2008. Visualization of the chemical space in drug discovery. *Comput. Aided Drug Des.* 4:322–33
63. Goktug AN, Chai SC, Chen T. 2013. Drug discovery. In *Pharmacology and Therapeutics*, ed. S Gowder, Chapter 7. Rijeka, Croatia: InTech
64. García-Domenech R, Gálvez J, de Julián-Ortiz JV, Pogliani L. 2008. Some new trends in chemical graph theory. *Chem. Rev.* 108:1127–69
65. Suh C, Sieg SC, Heying MJ, Oliver JH, Maier WF, Rajan K. 2009. Visualization of high-dimensional combinatorial catalysis data. *J. Comb. Chem.* 11:385–92
66. Awale M, van Deursen R, Reymond J-L. 2013. MQN-Mapplet: visualization of chemical space with interactive maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13. *J. Chem. Inform. Model.* 53:509–18
67. Klopmand G. 1992. Concepts and applications of molecular similarity, by Mark A. Johnson and Gerald M. Maggiora, eds., John Wiley & Sons, New York, 1990, 393 pp. Price: $65.00. *J. Comput. Chem.* 13:539–40
68. Willett P, Barnard J, Downs G. 1998. Chemical similarity searching. *J. Chem. Inform. Model.* 38:983–96
69. Chen X, Reynolds C. 2002. Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *J. Chem. Inform. Model.* 42:1407–14
70. Godden JW, Bajorath J. 2006. A distance function for retrieval of active molecules from complex chemical space representations. *J. Chem. Inform. Model.* 46:1094–97
71. Haranczyk M, Holliday J. 2008. Comparison of similarity coefficients for clustering and compound selection. *J. Chem. Inform. Model.* 48:498–508
72. Coifman RR, Lafon S. 2006. Diffusion maps. *Appl. Comput. Harmon. Anal.* 21:5–30
73. Platts J, Butina D, Abraham M, Hersey A. 1999. Estimation of molecular linear free energy relation descriptors using a group contribution approach. *J. Chem. Inform. Model.* 39:835–45
74. Liu ZK, Chen LQ, Rajan K. 2006. Linking length scales via materials informatics. *JOM* 58:42–50
75. Balabin RM, Lomakina EI. 2011. Support vector machine regression—an alternative to artificial neural networks for the analysis of quantum chemistry data? *Phys. Chem. Chem. Phys.* 13:11710
76. Balabin RM, Lomakina EI. 2009. Neural network approach to quantum-chemistry data: accurate prediction of density functional theory energies. *J. Chem. Phys.* 131:074104
77. Pilania G, Wang C, Jiang X, Rajasekaran S, Ramprasad R. 2013. Accelerating materials property predictions using machine learning. *Sci. Rep.* 3:2810
78. Rajan K, Suh C, Mendez PF. 2009. Principal component analysis and dimensional analysis as materials informatics tools to reduce dimensionality in materials science and engineering. *Stat. Anal. Data Min.* 1:361–71
79. Dewar MJS, Trinajstic N. 1969. Ground states of conjugated molecules—XIV. *Tetrahedron* 25:4529–34
80. Bajorath J. 2001. Selected concepts and investigations in compound classification molecular descriptor analysis, and virtual screening. *J. Chem. Inform. Model.* 41:233–45
81. Searls DB. 2005. Data integration: challenges for drug discovery. *Nat. Rev. Drug Discov.* 4:45–58