ELSEVIER

In pharmaceutical discovery, Hit-to-Lead strategies and processes are rapidly evolving and yet far from mature. A computational chemistry perspective can help projects deal with discovery risks and probabilities, and can help ensure effective decision-making.

Strategies and tactics for optimizing the Hit-to-Lead process and beyond—A computational chemistry perspective

Charles J. Manly, Jayaraman Chandrasekhar, Joseph W. Ochterski[1], Jack D. Hammer[2] and Benjamin B. Warfield

Neurogen Corporation, 35 N.E. Industrial Road, Branford, CT 06405, United States

The Hit-to-Lead-to-Candidate process continues to evolve rapidly, and while technological advances offer much potential, the reality often pales to the promise. Conversely, strategies and tactics implementing existing technologies may result in more benefit in the end. This article focuses on some of the thinking and approaches that may improve the efficiency and effectiveness of the beginnings of the drug discovery path. From the perspective of computational chemists, different types of strategy and philosophy of approach will be treated including: considerations of early lead choices, strategies for improving poor leads, multivariate optimization, opportunities for informatics, and engineering good decisions.

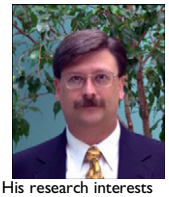
It is widely recognized that the drug discovery process is time-consuming, expensive, complex, and periodically impacted by paradigm-shifting technologies. Over the last few decades the process has evolved into one involving a series of loosely coupled and overlapping stages: Identification of a target → Generation of screening hits → Selection of lead(s) for optimization → Identification of a pre-clinical candidate by optimizing *in vitro* properties and efficacy in animal models → Develop- Development of a clinical candidate → trials to assess safety and efficacy in humans. Since we are dealing with a highly networked and complex series of biological processes with variability even at the level of the individual patient, attrition at every stage is inevitable. To avoid failure at later stages, which carries greater penalty in terms of resources, time and lost opportunity, it is extremely important to pay attention to the choices made and strategies employed at earlier stages [1].

In the Hit-to-Lead-to-Candidate process, computational chemists see some issues in the same way as other drug discovery scientists and may see some other issues differently. Often, we have backgrounds that include analysis of populations of compounds and data, focusing less on individual compounds and data points, and this may be different from other scientists.

Corresponding author: Manly, C.J. (cmanly@nrgn.com)

[1] Present address: East Hampton High School, 15 North Maple Street, East Hampton, CT 06424, USA. E-mail: ochtersk@yahoo.com.

[2] Present address: Rib-X Pharmaceuticals, Inc., 300 George Street, Suite 301, New Haven, CT 06511, USA. E-mail: jhammer@rib-x.com.



Dr Manly is VP, Discovery Technologies, and CIO for Neurogen Corporation. He received his Ph.D. in Chemistry from Harvard University following his B.S. in Chemistry and minor in Mathematics from The University of North Carolina. His research interests include the development of R&D processes to improve information and knowledge sharing and decision making, integrated high-throughput discovery, and the development of automated and assisted methodologies for pharmacophore recognition and model development. He has served on the advisory boards for several journals, conferences and institutes. Dr Manly has over 20 years of applied experience in the analysis and design of biologically active compounds with several companies and has been with Neurogen since 1993.

Dr Chandrasekhar is Associate Director, Biophysical Chemistry, at Neurogen Corporation. He received his Ph.D. from Indian Institute of Technology, Madras and subsequently carried out research in the groups of P.v.R. Schleyer in Erlangen, Germany and W.L. Jorgensen at Purdue University. Dr Chandrasekhar has served as a faculty in the Department of Organic Chemistry, Indian Institute of Science, Bangalore for 16 years. He has been with Neurogen since 2001. His current research is focused on using diverse modeling technologies to advance the drug discovery process.

Dr Ochterski earned his Ph.D. in Chemistry at Wesleyan University and completed postdoctoral studies at Yale University. After six years developing and supporting high performance computational chemistry software, he joined Neurogen as a computational chemist. There he developed novel modeling strategies in drug discovery and researched the effects of physical properties on the drug discovery and development processes. He is currently training the next generation of chemists as a secondary school chemistry teacher in East Hampton, Connecticut.

Dr Hammer obtained his Ph.D. in physical chemistry from Yale in 1999 and joined Neurogen as a computational chemist. From 2006, he has been working in the structure-based drug design group at Rib-X Pharmaceuticals to develop novel antibiotics that target the bacterial ribosome. He also continues to pursue his interests as a free-lance violinist, with credentials of being a former member of St. Louis Philharmonic and New Haven Symphony orchestras.

Benjamin Warfield has worked for Neurogen Corporation as a researcher and software developer since 2000. He received his B.S. in Chemistry from Yale University.

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2007.10.019

The authors posit that no one, including themselves, knows how to optimally pursue drug discovery, but it is felt that a variety of considered approaches and strategies make success more likely. In this article are highlighted some of the key issues the authors take into account in the Hit-to-Lead-to-Candidate process. We will start with discussions of some of the more important features of Hits and Leads and then move to exploration and optimization considerations. We also discuss some strategies that have worked well in the authors' hands in several discovery efforts involving varying challenges.

Key considerations

The importance of having a good starting point—'lead likeness' and ligand efficiency

The desirable characteristics of an oral drug candidate are easily specified: potent on-target efficacy, negligible off-target activity, and a reasonable ADME profile. In addition to *in vitro* potency, several physicochemical properties represent useful indicators of the drug-likeness of a compound. The importance of physicochemical properties, especially solubility and the dissolution profile, cannot be overemphasized. These have an impact not only on the ease of development of a drug compound, but on the extent to which it is absorbed, the fraction that reaches the desired site of action, and the manner and extent to which it is cleared from the system. Other physicochemical properties such as permeability, log *D* and polar surface area also play a key role in the 'druggability' of a compound.

It needs to be emphasized that the definition of solubility is not restricted to the equilibrium solubility of the drug. While this quantity is a key parameter for the formulation chosen for development, a more general assessment of solubility is needed [2–3]. The extent to which the drug (i) is available in solution at different pH states, (ii) manages to cross different membranes, (iii) is absorbed and (iv) is available in a free state to reach the target site of action are all to be factored into the concept of dissolution level of a compound. It is also important to understand the causes of poor equilibrium solubility in a compound, each of which brings its own challenges in formulation. For some highly lipophilic compounds, it will come as a result of very weak interactions with water, while for other high melting point compounds it will come as a result of efficient packing in the solid state, for example, through extensive hydrogen bonding. As will be discussed in a later section, the interpretation of *in vitro* data is also impacted by the effective solubility of a compound since the measurement of activity is directly controlled by the extent of sample available at the biological site of action in an assay.

There have been many analyses which profile the properties of compounds that have advanced to various stages of drug development [4–11]. Figure 1 shows the results of an in-house analysis of oral drugs launched during 1984–2002 ($N = 374$, anti-infectives excluded). The desirable ranges of molecular weight and lipophilicity are quite evident. These confirm the original Lipinski Rule of Five which has served as a particularly useful guideline in drug discovery. The odds of finding successful orally administered drug candidates drop significantly as molecular weight, lipophilicity, hydrogen bond donor and acceptor counts exceed key thresholds (Table 1). In order to find the ultimate drug candidate with the desired bounds in physical properties, it is necessary to start with

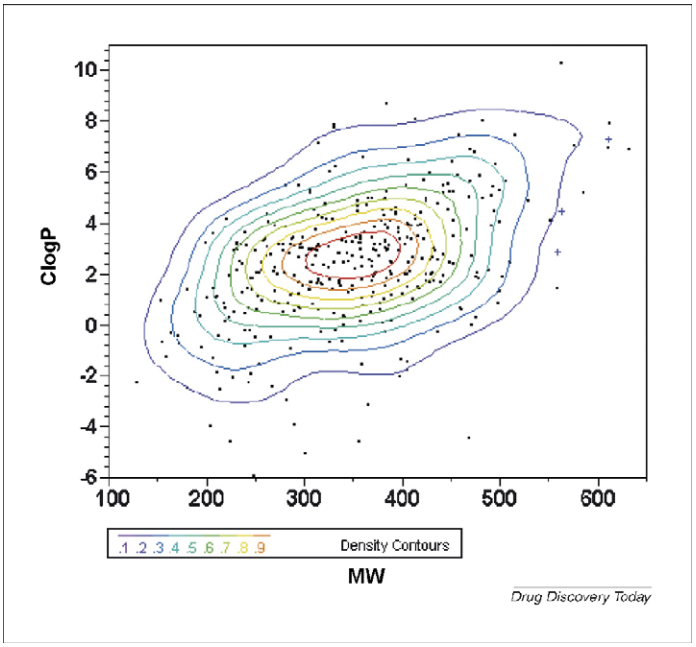


FIGURE 1
Distribution of molecular weights and Clog*P* of oral drugs launched during 1984–2002.

Leads with even tighter constraints. This is because the optimization process from lead to drug most often entails increases in molecular weight, lipophilicity, ring and rotatable bond counts (Table 2) [5,12,13]. Many of the improvements needed as the lead is converted to a drug candidate, such as reducing off-target interactions or improving metabolic stability, may be most easily achieved through the addition of functional groups that increase MW and often lipophilicity. Based on these studies, a set of thresholds for lead-likeness can be specified, for example, MW < 400 and log $P < 3.5$, to rank order hits for their suitability for further exploration and optimization.

Another convenient metric for assessing the attractiveness of a starting point is a quantity known as 'ligand efficiency'. One definition of ligand efficiency is potency in log units per heavy atom in the molecule [14]. Alternatively, it can be defined as pK_i/MW, a convenient measure of ligand efficiency that captures the spirit of the analyses of Lipinski and Oprea. This definition is one which we have found very useful, since MW is directly related to

TABLE 1
Key thresholds for physical properties

Property	Undesirable range	Percent of USAN[a] compounds over threshold [4]
Clog*P*	>5	10%
MW	>500	15%
HBD	>5	8%
HBA	>10	12%
MW + Clog*P*	>500 >5	1%
Polar surface area	>80 Å^2 >150 Å^2	Poor BBB penetration [40] Poor membrane permeability [10]

[a] United States Adopted Name.

TABLE 2

Average changes in physical properties from lead to drug [12]

Property	Change from leads to drugs
MW	+69
Clog*P*	+0.43
# Rings	+1
HBAs	+1
Rotatable bonds	+2

many of the properties we are trying to monitor and control, more so than heavy atom count.

One of the key goals of screening efforts should be to provide good starting points for further optimization. Next, we discuss the necessity to fine-tune the optimization process depending on the nature of the lead that is being optimized.

Exploration: where should we look, and what are we likely to find?

A common start for a project is to screen by various prioritization mechanisms until actives of a 'desired' level of activity are found and then spin off discovery efforts around these. This is a possible strategy as long as compounds of sufficient activity come from the screen early. However, this strategy is not necessarily optimal. Often, the early compounds that show most activity are not the best lead-like starting points for exploration and optimization [15]. There are thermodynamic arguments to suggest that the deck is even stacked in this direction; some of which will be discussed later.

An obvious place (though, as is highlighted later, not necessarily a recommended place) to begin looking for leads, from an ease of high-speed synthesis perspective, is in regions of relatively high lipophilicity and flexibility. Compounds from these regions are easier to synthesize since they tend to lack polar functional groups which can lead to undesirable side reactions and make generally applicable purification procedures difficult to define.

These compounds from a more flexible, lipophilic region of chemical space tend to result in higher hit rates than compounds from a more polar region, which at first consideration makes it seem like a more fruitful place to look for leads. However, one needs to be aware of the ramifications of searching in this chemical space. For a polar portion of a molecule to have a stronger interaction with a target than it does with water, it must interact with a polar region in the target. Polar–non-polar interactions are generally weaker than polar–polar interactions. Further, for maximum strength, the polar–polar interaction needs to have a fairly specific alignment, possibly along a hydrogen bond axis.

A non-polar portion of a molecule will interact weakly with water, but more strongly with a non-polar portion of a target; non-polar–non-polar interactions are usually stronger than polar–non-polar interactions. Moreover, non-polar interactions are not as sensitive to specific directionality in the same way as polar interactions. Therefore, lipophilic compounds can have many small interactions with the target, rather than a few strong ones, making it more likely that flexible lipophilic compounds will interact with a binding site and provide some indication of potency. See Figure 2 for a schematic explanation of these different types of interactions.

The authors feel that these non-specific interactions can make it more difficult to optimize such a compound, since the same characteristics also make it more likely that the compound will have off-target activity and other liabilities. In addition to pursuing these non-specific interactions, we suggest that a library built of smaller, more polar fragments in many orientations will yield more specific interactions and lower off-target activity. This strategy will undoubtedly result in fewer screening hits; however, those found will be of significantly higher quality as starting points.

There are approaches which can mitigate lower hit rates. Using a simple model of ligand–receptor interactions, a case has been made that the chances of finding specific interactions are low from a library constructed of complex molecules [16]. Because polar interactions are orientation sensitive, developing libraries which are locally uniform in terms of overall structure, but diverse in terms of polar group orientation is a good way to increase the likelihood that strong polar interactions will be found. (A carefully crafted library expansion strategy has been designed by the chemistry leadership at Neurogen and implemented through the efforts of high speed synthesis and medicinal chemists, taking

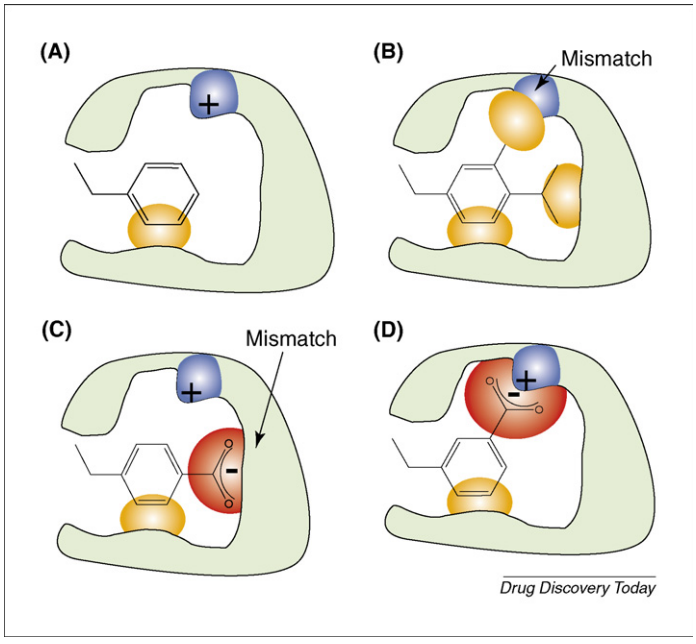


FIGURE 2

Schematic explanation of various types of interaction between compound and target. Diagram A shows a hypothetical partial interaction of an initial hit (the benzyl group) with the target (green). The yellow area indicates a non-polar interaction with the target. Diagram B shows the hit after further exploration—more areas of interaction will result in somewhat tighter binding and greater lipophilicity will favor binding in the target, increasing the apparent binding energy of this compound. Binding interactions with the target are non-specific, and include a mismatched non-polar–polar interaction. In diagram C, we show a more polar, non-optimized compound. This compound has a non-polar interaction with the target, and also has a mismatched polar–non-polar interaction. Since this compound will likely be more soluble than the compound in diagram A, the apparent potency will most likely decrease although the interaction with the target is stronger. Finally, in diagram D, the compound exhibits a strong polar–polar interaction as well as a non-polar interaction. This compound is likely to demonstrate higher potency and greater solubility with respect to the compound in diagram A. Further, since the directionality of polar interactions is more important than in non-polar interactions, the compound in diagram D probably will be more selective than that in diagram B.

into consideration effective probing for various interactions in an unmapped binding site, appropriate conformational flexibility, and lead-like starting points—Chenard, B.L., unpublished.) Fragment-based screening represents another attractive strategy along these lines [17–18]. Using flexible core structures will also permit more specific polar interactions to occur. Finally, screening at higher concentrations will make weaker leads more evident, and give a better indication of where polar interactions are to be found.

Lipophilic compounds have other potential liabilities as well. First, data interpretation can be more difficult for these compounds. The role of non-specific binding is particularly confounding with lipophilic compounds [19–21]. Second, if the compound is intended for use as an orally administered drug, the sparing solubility of lipophilic compounds can engender other risks as well. Low solubility can be associated with poor permeability, limited absorption and in turn, can contribute to inadequate bioavailability, one of the leading causes of drug failure in development [2]. The authors note that poor solubility and high lipophilicity are markers, although again not necessarily causative, for other risks as well, potentially including: limited exposure at site of action which leads to limited efficacy, often higher volumes of distribution, greater variability in oral dosing data (including but not limited to efficacy data), greater fed/fasted effects, slow onset of action, long t_{max}, adipose accumulation and retention, long terminal half-life, poor blood–brain barrier penetration, and increased form and formulation work to try to circumvent solubility problems [21]. In general, the decision to advance a limited solubility compound may imply a longer, more costly route to and through the clinic, with greater risk of failure along the way. This does not mean the compound cannot be a drug, but limited solubility means we must ask more questions, run more experiments, spend more thought interpreting the results, and understand that the risks are greater.

The pursuit of optimization and improving poor leads early

The most important factor necessary to efficiently optimize a lead is starting with the right lead in the first place. Many times a lead is chosen primarily because it is the most potent compound available. The authors prefer to not neglect the use of potent compounds which are smaller and more polar (i.e. with higher ligand efficiency). Hajduk has made a similar suggestion of taking into account both the size and potency within the context of fragment-based drug design [22]. Since optimizing a lead often increases its molecular weight and lipophilicity [12], starting with a better lead results in a more drug-like candidate in the end. In fact, the arguments outlined above regarding choosing smaller, more polar leads and avoiding more lipophilic compound suggest a reason why Oprea found that molecular weight and lipophilicity increase during lead optimization.

Sometimes, despite the best strategies in place for the early screening of hits, the lead we have on hand does not represent a good starting point: the molecular weight is too high or the compound is too lipophilic. In such cases, it is necessary to consciously work on improving the physical properties during the optimization, and, especially in the early stages of optimization, the improvement of properties is probably more important than the improvement of potency. One approach is to systematically remove parts of the compound to determine the relative importance of the different components of the lead. Another avenue of research along these lines is to move polar functionalities, especially hydrogen bond donors and acceptors around the compound, starting locally. This can potentially form better hydrogen bonds boosting the potency and selectivity of the lead. Introduction of conformational constraints or funneling the conformational population to the desired regions are standard medicinal chemistry optimization techniques. Through entropic control, such changes can boost potency without increasing molecular weight and lipophilicity.

Often the compounds comprising the best lead concepts are those which already have reasonable values for properties which are costly and difficult to evaluate. (This philosophy is central to the concept of 'lead-likeness'.) Optimization of any property (e.g. potency, efficacy, solubility, metabolism, etc.) proceeds best if a reasonable amount of data can be generated quickly to drive the optimization effort. A troubling situation arises when potency (for example, where data are usually readily generated) is easily optimized, but some other key property that must be improved for clinical candidate selection is costly or slow for data generation. In such a situation, the project team has difficulty understanding the SAR (structure activity relationship) around the property, and hence, must rely more on luck to optimize the parameter. On the other hand, a sub-optimal property in which data are easily generated often optimizes more quickly, since an SAR for this property is more easily generated.

It is not uncommon for lead optimization to follow a steepest decent, single direction at-a-time approach: find a highly potent compound, then try to increase its polarity to fix solubility problems, and so on. An alternative to this approach, multivariate optimization - simultaneous optimization of physical and pharmacological properties, is often more efficient. One means to this is the use of objective functions, which are essentially weighted averages of various properties combined into a single variable that scientists can follow. The components and their weighting may need to vary depending on the needs of the project—they are different from project to project, and often vary over time even within a project.

Potency typically has a larger weighting coefficient than other properties since it is a critical component of efficacy. However, it is possible for other properties in combination to outweigh potency leading to a less potent compound with better properties overall. It is important to be prepared to lose some activity in favor of addressing other problems, especially in early optimization of a lead with poorer properties.

Finally, one must be skeptical regarding any 'essential' or 'best' fragments of compounds which are conserved through the optimization lifetime. If these fragments are large or lipophilic, they can stand in the way of efficient optimization. There should consistently be some effort made to explore away from and replace those fragments—they pose a substantial risk to a project if a fatal flaw in that fragment is discovered further down the path.

Cost–benefit analysis of substituent changes—bioisosteres and beyond

Lead optimization typically requires rapidly understanding generalizable SAR for potency as well as for other properties, using well chosen substituent replacements at different parts of the molecule.

It is often tempting to retain all changes that lead to a boost in potency and eschew those that result in a loss of potency. A multivariate lead optimization strategy must balance the cost and benefits of each change from a broader perspective. The effect of substituent replacements on solubility, protein binding, membrane permeability, and so on. deserve to be taken into account while assessing the value of the replacement on potency. In particular, an increase in potency resulting from a large increase in molecular weight and/or lipophilicity has a built-in disadvantage that may have serious repercussions later in development.

An obvious focus of interest has been identifying bioisosteric replacements [23–24]. Much of the knowledge of structural changes that do not significantly alter biological activity has been accumulated through medicinal chemistry optimization efforts. Attempts have been made to codify the in-house corporate knowledge, such as specific core or substituent replacements known to individual chemists or project teams, into software tools that are amenable to all chemists cutting across projects [25]. A more general set of functional group replacements that serve as effective bioisosteres for specific targets can be identified through computational analyses, especially when a large quantity of data is available. These would offer novel ideas not often obvious even to an experienced medicinal chemist. Neurogen has captured and encoded bioisosteric information and integrated it into the NDL programming language described below. Since this information includes successful, non-obvious replacements from individual projects, it can provide useful suggestions for chemists in other projects.

Interpretation of the value of a non-bioisosteric functional group replacement is a critically important component of lead optimization. There have been several studies carried out on the effect of different common substituent replacements on a variety of properties of relevance to lead optimization [26–28]. For example, Haubertin and Bruneau [27] have organized the effect of 0.7 million substituent replacements involving 9000 different functional groups on a variety of experimental properties. They have described an application in which chemists can draw on the historically observed trends to assess ideas to improve solubility, protein binding and log *D*.

At Neurogen, the authors have carried out analyses of the effects of functional group changes on a variety of biological activity data. As pointed out earlier, increases in lipophilicity generally increased the median potency across multiple GPCR and ion-channel targets. It was possible to develop a set of guidelines for identifying the likely origins of potency change resulting from the substituent replacement. In particular, a specific interaction with the target is implicated only when the potency boost was above a certain threshold expected on the basis of the more non-specifically seen change in lipophilicity. Conversely, hints of a specific interaction with a polar group could be surmised even when the overall effect was a slight reduction in potency, as long as the reduction was less than that anticipated on the basis of the lipophilicity change alone (Figure 3).

This analysis also offers an attractive framework for quantitatively assessing the replacements worth retaining and those that, in return, involve too high a penalty in the drug-likeness of the resulting compounds.

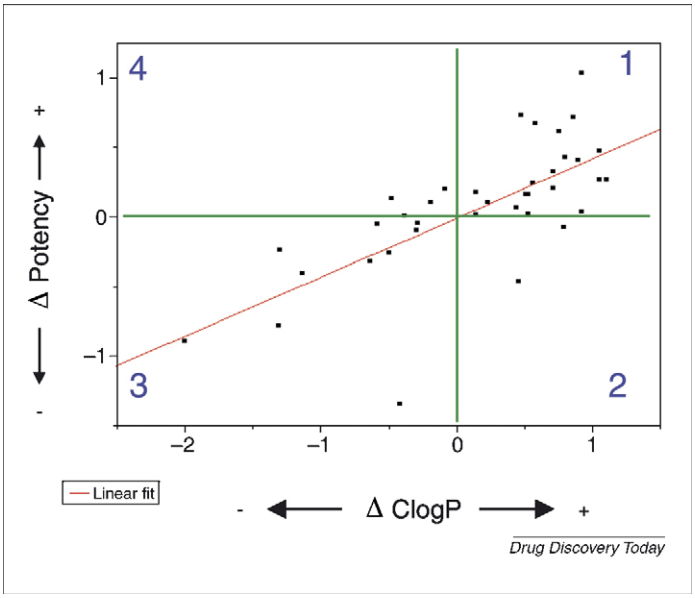


FIGURE 3
Representative plot of the effect of different substitution replacements on potency and lipophilicity. A top level analysis is as follows: If the change increases the lipophilicity, the potency should have improved by more than '*X*'-fold as determined from the plot (quadrant 1). The expected value represents just the median, and implies no significant specific molecular recognition in the region of the substituent change. If the potency decreases as in quadrant 2 (or increases less than '*X*'-fold), there is presumably a potential steric problem with the compound at the binding site or an important conformational change is involved. If the substituent replacement decreases lipophilicity, a reduction in potency is to be expected (quadrant 3). If the reduction in potency is less than the median amount from the plot for the substituent change, there is potentially a specific molecular recognition interaction to be made with the binding site. There should have been a loss of '*X*'-fold potency due purely to lowered energy of the solvated state. Less than '*X*'-fold loss or any gain (quadrant 4) in potency represents an attractive opportunity warranting further exploration of another such group or moving the group one position over or one atom out on the tether to achieve more optimal interaction.

Strategies

It is not a safe approach to pursue a single lead, however attractive it may appear at any given point. The potential for a new problem at the next stage of optimization should never be overlooked. While dilution of effort may be a concern to some, it has been the preferred approach at Neurogen to simultaneously optimize more than one lead. Frequently, the differences in SAR patterns, especially for potency, turn out to be parts of a larger and more inclusive SAR, and the strategy of pursuing SAR this way has been quite beneficial. Often, an insight derived from one lead concept, for example, a substitution pattern that reduces hERG liability, proves to be an advance for other leads as well.

It may appear that pursuing multiple leads may cause confusion rather than provide clarity during optimization. We prefer to use the term 'exploring multiple lead concepts'. Before proceeding, let us discuss this idea of 'concept' a bit more.

Concepts to compounds

Concepts include, but are not the same as, compound series. 'Series' has specific meanings to medicinal chemists and others on a project, often defined largely by project history. Concept includes series but may represent subsets of series, such as under-

sampled subsets which should be examined further, or supersets of series which combine portions of multiple series through the same pharmacophore. As one example, two compounds with differing heterocyclic 'cores' might normally be considered two different series, but from the point of view of the binding site, if the spatial occupancy, molecular shape, and electrostatic, lipophilic and hydrogen bonding interactions are similar, these two may represent the same concept. The reason for focusing on concept is to examine and explore more meaningful groupings of compounds such as 'compounds adhering to a specific SAR', or 'compounds making use of a specific set of binding site interactions', or 'compounds that show efficacy in this model', etc. This focus may be more relevant to the drug discovery and development process than some other more arbitrary series definitions defined by similar 'core' or substituent chemistry functionality [29].

We would encourage project teams to spend a reasonable effort, at the beginning, identifying, generating and evaluating multiple lead concepts before focusing attention too early on a very limited number of concepts. Traditionally, in discovery research, projects have used primarily potency to focus efforts early on a very few lead series without knowing the liabilities of that series or how easily optimized these liabilities might be. By canvassing multiple concepts and evaluating each for strengths and weaknesses (by evaluating a few representatives from each), the hope is that better decisions for focusing resources can be made. In reality, some (perhaps low) level of effort should be dedicated toward these ends throughout much of the project lifetime.

Therefore, we have tried to adopt an early project strategy of 'fast and cheap to concepts, understand the positives and negatives of each, and then focus resources'. A highly integrated discovery platform which tightly couples, through informatics, the disciplines of high-throughput pharmacology, high-speed synthesis and computational chemistry provides a strategic opportunity in the early stages of projects in this regard. This integration can easily take advantage of computational models to data mine and then physically mine (through directed HTS assay activities) active concepts. We have described this as 'directed sequential screening' or 'targeted sequential screening' ([30], see also [31]).

Similarly, the information space around these concepts can be mined giving a reasonable initial SAR for the concept. Then, molecular modeling predictions and high-throughput property assays can describe the drug-like property space for the concept, thus bringing together significant information about activity, 'developability', etc. for each concept so that positives and negatives of each can begin to be assessed [30].

Informatics and data mining

One technique which has allowed us to quickly find higher quality leads is to screen pre-selected portions of our archives up front. In particular, one of the first sets of compounds to go through primary screening is a pre-identified set of lead-like compounds, chosen for low molecular weight and greater hydrophilic character. Hits from this part of the archive are certainly of higher value, but are also less likely to occur (see above discussion). To offset this risk, we simultaneously prioritize other portions of our archive for screening as well. For example, we have designed a set of compounds which reflects the diversity of our archive, and a set of racks which emphasizes conserved binding motifs (or privileged templates) from compounds active at more than one target [32–34]. Finally, we also incorporate compounds which are similar in structure to known active compounds, as well as compounds predicted by *in silico* models to be more active. All of our screening efforts are comprised a blend of these streams, balanced depending on the current needs of each project.

The technique also extends to *in silico* screening of our virtual library. Here we have carved out a special section of the virtual library which has several desirable properties: low molecular weight, low lipophilicity, relatively high solubility, and high likelihood of synthetic success and immediate availability of synthesis components. This portion of the library, which we call PriVLib (Privileged Virtual Library), has been a fruitful source of new leads for us.

The PriVLib is constructed in the following manner. First, a virtual library is formed from combinations of synthesis fragments we have in inventory and robust synthesis protocols. By robust, we mean that the given protocol has a very high likelihood of synthetic success in terms of producing useful quantities of the correct purified product using robotic synthesis techniques [30,35]. We then filter this virtual library to obtain only low molecular weight, low lipophilicity compounds. Exactly which values are used for these and all other criteria depends on the needs of each individual discovery project, but typically the criteria are chosen to be more lead-like than merely drug-like. Finally, we predict (using *in silico* models) solubility and activity for the resulting compounds. If necessary, we can also filter for other desirable predicted properties such as microsomal stability or reduced hERG channel activity. After human inspection, these compounds are prioritized for synthesis.

Because this privileged portion of our virtual library is two orders of magnitude smaller than our full virtual library, and filtering reduces that still further, we can use our highest quality, most computationally expensive models to give us the greatest confidence in our selection of the final list of compounds to generate.

Neurogen data language

In order to successfully choose the best concepts to pursue, it is necessary to be able to distinguish one concept from another. This in turn, again, raises the question: what is a 'concept'? One way to think of it is as a series, but our use of the term concept goes far beyond that; we mean it in the most general sense possible. A concept could be compounds with a similar substructure within a certain range of molecular weight and lipophilicity. It could be compounds from several series, all of which have the same overall shape, say rod-like or Y-shaped, within a certain potency range. In reality, a concept can be any grouping of compounds that a scientist can perceive that may result in actionable decisions. This generality leads to a difficulty in terms of using concepts practically: namely, how can a scientist gather a list of compounds in a concept without wading through the (possibly hundreds of thousands of) structures and properties manually?

To address this problem, we have developed a computer language (NDL, which stands for Neurogen Data Language) to efficiently categorize compounds into concepts. Although this language is certainly used routinely by scientists to classify compounds into concepts and to compare those concepts to each other, its

```
1) gather_table XYZ_Data of samples where project is ('XYZ') since 1/1/2007;
2) filter in substructure matches "c1ccc2nccc2c1";
3) filter out MW > 500;
4) Concepts = create_groups first_match
5)    Phenols => substructure matches "Oc1ccccc1",
6)    LWM_Thiophenes => substructure matches "c1ccsc1" and MW < 350,
7)    Unstable => Predicted_Rat_Microsomes < 50,
8)    otherwise Others;
9) dsa;
10) OBJ_FN = (Predicted_Solubility > 75) + (MW < 450 ) + (CLogP < 3)+ 3*( Potency > 8 );
11) sort OBJ_FN descending;
```

FIGURE 4
A simulated NDL (Neurogen Data Language) script illustrating several capabilities of the language. Line 1 shows a typical statement for gathering together a starting list of compounds. Similar statements can be used to gather groups of compounds having other criteria in common: substructure, similarity to some probe compound, synthesizing chemist, or run in the same assay. There are also several ways to define a range of time to search through. Lines 2 and 3 illustrate that compounds can be filtered in or out depending on arbitrary criteria—substructure and molecular weight (MW) in the example. Data lookups, calculations and predictions occur on-the-fly for both real and virtual compounds when needed. Lines 4 through 8 demonstrate the flexibility of NDL's ability to define concepts. Each compound is compared against the criteria of the concept definition, and assigned the label (to the left of the arrow '=>') of the first criteria matched. If none of the criteria fit, compounds can be deleted from the list or assigned an arbitrary label ('Others' in the example). It is also possible to assign multiple labels to a given compound. In line 9, we show that NDL is tightly coupled to our informatics tools. The 'dsa' statements invokes our proprietary 'distinct series analysis' tool. Objective functions can be constructed from complex expressions involving any available data source. This is illustrated in line 10. Finally line 11 shows that NDL can sort the final list by any criteria desired.

applications have grown to address a much wider range of questions. It is used to organize and present data to project team members in ways allowing them to see subsets of compounds important to a project: the most recent compounds, the most potent compounds, the compounds with the best multivariate properties and the compounds which have received the most attention. NDL is designed to make it easy to generate subsets of compounds based on their synthesis date, results in an assay, date they were run through an assay and the like. For example, it is easy to ask for all compounds which were run through assay × during the last 6 months which fell into a certain activity range. A sample of simulated NDL code is given in Figure 4. Part of the output resulting from running the NDL script is shown in Figure 5 which describes the sequential execution as it progresses. In a real example, this information would be followed by structures along with the requested experimental and predicted data grouped by distinct chemical series.

Along with the compound structure our designed views typically incorporate the most important data about the compounds, including molecular weight, lipophilicity as well as the results of a battery of experimentally generated data. Because of the tight integration of our data storage structure and computational systems, we can freely mix experimental, calculated and predicted data into a single view and combine them as desired in objective functions.

NDL is coupled to a custom-built web-based data display system. This system allows data sets to be redisplayed in whatever format seems most informative to the current user, as individual data points (with or without structures), subsets, or summarized by 'concept' (with histograms showing the distributions of important properties) as discussed above. In the event that the in-browser display and sorting capabilities provided are insufficient, the data set can be instantly exported into a third-party data analysis tool, or returned to NDL for further augmentation or filtering. Other features built into NDL are an automatic structure–activity relationship (SAR) table generator, a bioisostere generator as well as several tools which find pairs of compounds with specific relationships, such as substructure replacements.

Other learnings regarding exploration and optimization

The meaning of K_i and assessment of binding

In simplest terms, the inhibition constant, K_i, is generally assumed to be the difference in free energy between the bound and unbound complexes. However, it is not uncommon for variations in K_i to be interpreted in terms of only the interactions in the bound complex, leaving out an important aspect of understanding these changes. Naturally, all experimental systems involve non-specific binding which complicates how we interpret K_i and must be either corrected for or disregarded. If we simplify interpretation by disregarding non-specific binding, the resulting K_i is a measure of the free energy difference between the solvated compound and the compound that is bound to the target. If two compounds interact with the same energy to a target, the compound that is less effectively hydrated (typically the more lipophilic compound) will yield a lower K_i with higher apparent potency. While it is difficult to individually measure relative free energies of hydration and free energies of interactions with a target receptor/enzyme, such decompositions have been carried out in computer simulation studies of relative free energies of binding (Udier-Blagovic, M. and Jorgensen, W.L., unpublished) [36–37]. Higher apparent potency involving lipophilic compounds is more often because of the reduced desolvation penalty, rather than stronger interactions with the target. Thus, more polar and generally soluble compounds are at a disadvantage since they may appear less potent although the strength of interactions with the target may actually be greater. When choosing a lead concept or path of optimization, this suggests that more polar compounds – even with less apparent potency – should be examined further.

```
start gather_table XYZ_Data
end gather_table XYZ_Data:
   Nselected:7948 total_size:7948 Ncols:3
start structure_matches c1ccc2nccc2c1
end structure_matches:
   Nselected:7948 total_size:7948 Ncols:3
start filtering in
end filter :
   Nselected:1060 total_size:7948 Ncols:3
start creating column MW for 1060 IDs
end creating column MW:
   Nselected:1060 total_size:7948 Ncols:4
start filtering out
end filter :
   Nselected:1039 total_size:7948 Ncols:4
start creating column latest Predicted_Rat_Microsomes for 1039 IDs
end creating column latest Predicted_Rat_Microsomes:
   Nselected:1039 total_size:1039 Ncols:5
start structure_matches Oc1ccccc1
end structure_matches:
   Nselected:1039 total_size:1039 Ncols:5
start structure_matches c1ccsc1
end structure_matches:
   Nselected:1039 total_size:1039 Ncols:5
start dsa similarity_threshold 0.6 desctype qfyd
end dsa:
   Nselected:1039 total_size:1039 Ncols:10
start creating column Potency for 1039 IDs
end creating column Potency:
   Nselected:1039 total_size:1039 Ncols:11
start creating column Predicted_Solubility for 1039 IDs
end creating column Predicted_Solubility:
   Nselected:1039 total_size:1039 Ncols:13
start creating column CLogP for 1039 IDs
end creating column CLogP:
   Nselected:1039 total_size:1039 Ncols:14
start sorting NDL_sort descending
NDL done(1039 data points left).
```

FIGURE 5
Part of the output resulting from running the NDL script shown in Figure 4.

Data for insoluble (sparingly soluble) compounds

When a compound is less soluble than its nominal concentration in an assay, extra care should be taken while interpreting the results of the assay. For example, if the nominal assay concentration is 1 μM (1000 nM), but the solubility of the compound is 50 nM, then the true effect of the compound at 1 μM cannot be seen. If the result suggests that the compound is inactive, one cannot be certain about its significance—the compound may be inactive, or it may be active at some concentration higher than 50 nM. On the other hand, if the compound is active, then that compound may be much more potent than it appears, since the actual concentration may be much lower than the nominal concentration [21].

It is not an uncommon belief that a well-behaved (sigmoidal) dose–response curve is an assurance that data are reliable. This is only partially true. It is true that the compound is exhibiting a normal relationship between concentration and effect; however, for sparingly soluble compounds, the concentration of the compound will be lower (possibly orders of magnitude lower) than expected. Sparingly soluble compounds can be sequestered in proteins, cell membranes, along test tube or micro-array plate walls, and may not necessarily 'crash out' of solution. As long as the main sequestration channel is not saturated, the compound will partition uniformly. If, say the compound partitions 99% to the walls of the container and 1% into solution, then a normal dose response curve can result, albeit at 100-fold lower concentration [21].

This is true for both on-target and off-target activity. Since side effects are typically tested at much higher concentrations than on-target activity, the consequences of misinterpreting the results are especially onerous. Side effects may occur at much lower concentrations, removing the safety margin of a perceived therapeutic index.

In addition to the above points, the authors also acknowledge the effect of aggregates on assay results [38], and the varying effects of precipitates in assays.

Synthesizable workspace

It is important to keep synthesis moving forward in several concepts simultaneously. Naturally, the lead concept will be one, but we recommend also having an active concept with easy synthesis. Often, lead concepts have more difficult synthesis; therefore it is difficult to explore a wide range of replacements for various functional groups or the core. A less desirable concept which can be synthesized allows a much wider range of exploration to be performed—more ideas can be tested in this concept, and the most useful of those transferred to the lead concept(s).

Strategy and philosophy of implementation—decision engineering for impact

We all have seen, and developed, informatics systems which provided much of the information that a person, team or company needed, but missed some key piece of information or did not organize the data quite right. The result was that the application was used, and some found it useful, but the application never quite had the impact that we had envisioned and desired. This is especially troublesome when we were trying to provide information critical to an important decision.

We are always looking for ways to have more impact for our resource investment. One response is a structure which we call Decision Engineering. In Decision Engineering, we try to address specific decisions or the needs of other similar actions, and we try to provide all of the information needed for that decision in the organization of the output of the application. Part of the information has been requested by the potential users, but a significant part of the information is there because, even if not requested, it should be a key part of the decision making process. An effective decision requires that this information be taken into account, so it is integrated into the organization of the data. In this way, we are trying to engineer a good decision. We speak of it as 'providing the right information, organized the right way, at the right time, to make the right decision'.

One outcome of this philosophy is that many data appear in multiple forms on multiple web pages. While one might see this as needless overhead to be maintained, the same data may be valuable in multiple decisions and analyses, but perhaps should be organized differently for each. One example is that each different decision or need may require the same data to be interpreted differently—like a solubility value that is acceptable for one purpose, but not for another.

So, instead of simply tallying information, or providing only the information that has been requested, we try to understand the need beyond the request, and not just the request. We try to become partners in the thinking and decision making strategy, and we find that the systems that result usually have more of the impact we were striving for. This places an obligation on us, not just only to be good informatics scientists, but also to understand the specific science, technology, and decision making process well enough that we can contribute intellectually to the process, and this is often a high hurdle.

Part of the added information is output from simulation and modeling, which is calculated and added on the fly or retrieved from the data warehouse. For example, if solubility data are not yet available, but this would be important to the decision, then we provide a model prediction together with an estimate of the prediction quality.

One example of such decisions is progression of compounds in a project through the various stages, generating data and collecting information on them along the way. Generally this results in a filtering process whereby, compounds stack up at each assay on the Progression Flowchart, until data are generated, and some compounds advance to the next assay 'gate' and some don't make the 'cut' and are not progressed.

Project flowcharts should only be guidelines—hard and fast progression gates do not make for efficient or good decisions. Actually, what we want to do is find the most important information to make a decision on each compound and factor this into the prioritization of progression activities. It makes little sense to spend resource early in a flowchart generating data on a compound when other data, not these data, are key to its ultimate value. Each compound has its own key questions that are critical with respect to its most efficient advancement—it is largely an individual decision, compound by compound. This has to be factored and tempered with the need for efficiency in low-throughput and costly assays.

Social networking in knowledge management, decision engineering and idea evolution

We include below a couple of experiments in which we have attempted to improve decision-making and idea generation through information flow.

Project blogs

Web logs (blogs) can be a very useful technique for advancing and tracking projects and fostering knowledge sharing across projects. The impact is very dependent on the specific users and their personalities and the personality of the team, but blogs are inexpensive to develop and deploy, and sometimes they are used. Readily available open source solutions now exist to deal with the melding of text and images, such as structures, on the web (e.g. http://www.livejournal.com).

BEAGLE—idea sharing and idea evolution

In an attempt to promote the evolution and sharing of ideas within a project team, we developed Beagle (named for Darwin's ship). Beagle shares some functionality with blogs, but goes further, attempting to integrate modeling/simulation and fast retrieval of existing data on similar compounds from the archive with feedback from the expert community of a project. This provides individual users with the opportunity to validate or refute their own ideas based on predictions from models and on feedback pertaining to known compounds. They can evolve their ideas based on the ideas they see from others. They can benefit from feedback from the expert community on their ideas as they are evolving them, similar to an electronic brainstorming session over time.

Beagle is organized by project—combining assays, models, properties, objective functions important to the specific project. While the viewers/participants are experts in the project, they see things differently and see different opportunities and different paths forward from the same data and ideas. Beagle tries to help scientists evolve their ideas through this sharing and feedback.

The key functionality that Beagle tries to facilitate includes:

- Capturing ideas with some notation of what the creator was thinking.
  - Ease of entering structures and comments.
- Organizing structures into different types of concepts—easy, flexible browsing.
- Predicting activity and properties for molecules based on computational models.
- Finding the closest molecule in the compound archive.
  - Prioritizing it for an assay, if data do not already exist.
  - Examining the compounds and existing data.
- Sharing with others on the project.
- Seeing feedback from others.
- Seeing others' ideas with their thinking and rationale.

Beagle has a simple web interface, designed to work smoothly with traditional ideas of close SAR exploration. If a scientist visits this page wanting to explore new ideas, they will find a structure drawing tool into which a starting structure may be drawn by hand (or pasted in from an external application) or loaded from the company's structure database. The user can then modify this structure freely in the drawing tool, according to the user's own ideas of reasonable bioisosteric replacement and SAR exploration. An editable list of entered structures is maintained in the same window, to which the current structure in the drawing tool may be added with the click of a single button.

These structures and their calculated properties are then displayed in a sortable grid in the browser. If the results are interesting, users may decide to 'publish' them to be shared with the rest of the team, or to download them into a spreadsheet for further analysis. Alternatively, they may want to attempt further exploration around one of the structures just entered: links are provided on each row to return to the initial input page, loading that row's structure into the drawing tool as a starting point. This allows for rapid and cheap testing of ideas which might otherwise be lost in the background.

After the scientist enters an idea (compound), model predictions of physical properties and biological assays are immediately generated with estimates of prediction quality. The scientist can also easily see how close this compound is to others in the compound archive and which, if any, already have data generated.

Using the same interface (toggling between input and search modes is a simple in-browser buttonpress), users can also readily find other compounds similar to their own idea that others on the team have already considered. Seeing these may prompt them to take their idea forward in a different direction, leading to another idea compound and predictions for it.

Beagle may also provide information for forming the Intellectual Property strategy for the project, since it captures, in a way, the breadth of opportunity.

From a social networking perspective, Beagle incorporates several aspects of 'prediction markets' [39]. Prediction markets are characterized by an expert (and semi-expert) community voting on possible outcomes by making investments similar to the stock market. Beagle tracks 'drill-down' viewing of ideas and provides this as feedback, and in this way is similar to people voting in a prediction market.

Outlook and conclusions

While progress has been made, there is still much improvement to be gained in drug discovery from further integration of processes, data and simulation, together with better handling of knowledge streams to promote effective decision making. In the future, there will be better integration of experimental data and simulation/prediction in assessments and decisions. As an industry we tend to emphasize new and better data (new assays, better models, better predictions) rather than better usage of existing data and predictions we currently have.

Risks and uncertainties will be (and needs to be) better treated in decision making. As an industry, we tend to wish for more clarity and more certainty, but in many cases, increased certainty will be quite slow and costly.

What do we know? What do we not know? What do we think we know, but really do not know? Drug discovery and the Hit-to-Lead process will need to deal with uncertainties better, and learning how to deal effectively with these uncertainties and risks in decision-making will be an increased focus going forward.

We have tried to describe some of the directions our thinking has evolved regarding how to pursue the goals of Hit-to-Lead and beyond. Much of what has been described is not new or unique, and all of it has benefited greatly from the community of scientists pursuing these same endeavors and who have shared their thoughts. We appreciate the opportunity to share back in kind.

Acknowledgements

We thank Bertrand Chenard, James Krause and Stephen Uden for stimulating interactions and suggestions.

References

1 Lipper, R.A. (1999) E pluribus product. *Mod. Drug Discov.* 2, 55–60
2 Hendriksen, B.A. *et al.* (2003) The composite solubility versus pH profile and its role in intestinal absorption prediction. *AAPS Pharm. Sci.* 5, E4
3 Wang, J. *et al.* (2007) Development of reliable aqueous solubility models and their application in druglike analysis. *J. Chem. Inf. Model.* 10.1021/ci700096r (electronic pre-print article)
4 Lipinski, C.A. *et al.* (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 46, 3–26
5 Wenlock, M.C. *et al.* (2003) A comparison of physiochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* 46, 1250–1256
6 Vieth, M. *et al.* (2004) Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.* 47, 224–232
7 Vieth, M. and Sutherland, J.J. (2006) Dependence of molecular properties on proteomic family for marketed oral drugs. *J. Med. Chem.* 49, 3451–3453
8 Leeson, P.D. and Davis, A.M. (2004) Time-related differences in the physical property profiles of oral drugs. *J. Med. Chem.* 47, 6338–6348
9 Egan, W.J. *et al.* (2000) Prediction of drug absorption using multivariate statistics. *J. Med. Chem.* 43, 3867–3877
10 Kelder, J. *et al.* (1999) Polar molecular surface area as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.* 16, 1514–1519
11 Veber, D.F. *et al.* (2002) Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* 45, 2615–2623
12 Oprea, T.I. *et al.* (2001) Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* 41, 1308–1315
13 Oprea, T.I. (2002) Current trends in lead discovery: Are we looking for the appropriate properties? *J. Comput. Aided Mol. Des.* 16, 325–334
14 Hopkins, A.L. *et al.* (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* 9, 430–431
15 Rishton, G.M. (2003) Nonleadlikeness and leadlikeness in biochemical screening. *Drug Discov. Today* 8, 86–96
16 Hann, M.M. *et al.* (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* 41, 856–864
17 Fattori, D. (2004) Molecular recognition: the fragment approach in lead generation. *Drug Discov. Today* 9, 229–238
18 Erlanson, D.A. *et al.* (2004) Fragment-based drug discovery. *J. Med. Chem.* 47, 3463–3482
19 Margolis, J.M. and Obach, R.S. (2003) Impact of nonspecific binding to microsomes and phospholipid on the inhibition of cytochrome P4502D6: implications for relating *in vitro* inhibition data to *in vivo* drug interactions. *Drug Metab. Dispos.* 31, 606–611
20 Austin, R.P. *et al.* (2002) The influence of nonspecific microsomal binding on apparent intrinsic clearance, and its prediction from physicochemical properties. *Drug Metab. Dispos.* 30, 1497–1503
21 DeWitte, R.S. (2006) Avoiding physicochemical artifacts in early ADMET-Tox experiments. *Drug Discov. Today* 11, 855–859
22 Hajduk, P.J. (2006) Fragment-based drug design: How big is too big? *J. Med. Chem.* 49, 6972–6976
23 Sheridan, R.P. (2002) The most common chemical replacements in druglike compounds. *J. Chem. Inf. Comput. Sci.* 42, 103–108
24 Ertl, P. (2003) Cheminformatics analysis of organic substituents: Identification of the most common substituents, calculation of substituent properties, and automatic identification of druglike bioisosteric groups. *J. Chem. Inf. Comput. Sci.* 43, 374–380
25 Stewart, K.D. *et al.* (2006) Drug Guru: A computer software program for drug design using medicinal chemistry rules. *Bioorg. Med. Chem.* 14, 7011–7022
26 Lewis, R.A. (2005) A general method for exploiting QSAR models in lead optimization. *J. Med. Chem.* 48, 1638–1648
27 Haubertin, D.Y. and Bruneau, P. (2007) A database of historically-observed chemical replacements. *J. Chem. Inf. Model.* 10.1021/ci600395u (electronic preprint article)
28 Lewell, X.Q. *et al.* (2003) Drug rings database with web interface. A tool for identifying alternative chemical rings in lead discovery programs. *J. Med. Chem.* 46, 3257–3274
29 Lajiness, M.S. *et al.* (2004) Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem.* 47, 4891–4896
30 Manly, C.J. *et al.* (2001) The impact of informatics and computational chemistry on synthesis and screening. *Drug Discov. Today* 6, 1101–1110
31 Schreyer, S.K. *et al.* (2004) Data shaving: A focused screening approach. *J. Chem. Inf. Comput. Sci.* 44, 470–479
32 Bondensgaard, K. *et al.* (2004) Recognition of privileged structures by G-Protein Coupled receptors. *J. Med. Chem.* 47, 888–899
33 Bemis, G.W. and Murcko, M.A. (1996) The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* 39, 2887–2893
34 Bemis, G.W. and Murcko, M.A. (1999) Properties of known drugs. 2. Side chains. *J. Med. Chem.* 42, 5095–5099
35 Noonan, J.N. *et al.* (2003) Advancing parallel solution phase library synthesis through efficient purification. Quantitation and characterization techniques. *J. Assoc. Lab. Autom.* 8, 65–71
36 Saito, M. *et al.* (2005) A free energy calculation study of the effect of H->F substitution on binding affinity in ligand–antibody interactions. *J. Comput. Chem.* 26, 272–282

37 Sims, P.A. *et al.* (2005) Relative contributions of desolvation, inter- and intramolecular interactions to binding affinity in protein kinase systems. *J. Comput. Chem.* 26, 668–681
38 Shoichet, B. *et al.* (2002) A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem* 45, 1712–1722
39 Wolfers, J. and Zitzewitz, E. (2004) Prediction markets. *J. Econ. Perspect.* 18, 107–126
40 Clark, D.E. (2005) Computational prediction of blood–brain barrier permeation. In *Annu. Rep. Med. Chem. (Vol. 40)* (Doherty, A.M., ed.), pp. 403–415, Elsevier