4830 J. Med. Chem. 2010, 53, 4830–4841
DOI: 10.1021/jm100164z

Journal of Medicinal Chemistry
Miniperspective

Cheminformatic Tools for Medicinal Chemists

Steven W. Muchmore, Jeremy J. Edmunds, Kent D. Stewart, and Philip J. Hajduk*

Pharmaceutical Discovery Division, Abbott Laboratories, Abbott Park, Illinois 60064

Received February 5, 2010

Introduction

Cheminformatics can be broadly described as any attempt to use chemical information to infer the relationships between or attributes of chemical structures. From a drug discovery perspective, cheminformatic principles can be applied from the earliest stages of lead discovery (e.g., chemical similarity and library design) to lead optimization (e.g., QSAR studies) through to preclinical and clinical development (e.g., predictive toxicology). The popularity of cheminformatics and its use in academia and the pharmaceutical industry can be appreciated from the fact that at least five scientific journals exist almost exclusively dedicated to the field (*The Journal of Cheminformatics*, *The Journal of Chemical Information and Modeling*, *The Journal of Computer-Aided Molecular Design*, *Molecular Bioinformatics*, and *QSAR and Combinatorial Science*), and more than 15000 scientific journal articles have been published during just the last 5 years that describe cheminformatic research. This intense interest in cheminformatics stems from the promise that, if underlying relationships between a given chemical structure and a host of biological end points exist and can be elucidated, drug discovery timelines can be significantly reduced. Given the pressure on the pharmaceutical industry to increase productivity while decreasing costs, prior knowledge of which molecules have the highest probability of success (or at least knowing which molecules are likely to fail) is worthy of vigorous pursuit.

Over the past decade there have been several significant advancements in our understanding and application of cheminformatic principles. Approaches to measuring and comparing chemical information have become both more sophisticated and accessible. For example, two of the most powerful chemical similarity measures (two-dimensional (2D) extended connectivity fingerprints and three-dimensional (3D) shape and electrostatic overlays) are available in user-friendly software packages from Scitegic (Accelrys) and Openeye Scientific Software. Multiple methods for understanding and predicting bioactivity have proven their robustness, including partial least-squares (PLS), genetic algorithms, Bayesian analyses, and Random Forest analyses. Our understanding of molecular features or properties associated with certain pharmacological end points has also dramatically increased. For example, it has been widely recognized that certain structural features can be associated with toxicity, while other molecular properties (such as ClogP, molecular weight, and polar surface area) can be associated with oral bioavailability[1] or promiscuity.[2] Thus, the modern medicinal chemist has access, either directly or indirectly, to an enormous array of tools and methods (for examples, see refs 3−7) for improving the probability that newly designed molecules are potent, safe, and orally bioavailable.

Unfortunately, despite these and other dramatic advancements in cheminformatics research, the pharmaceutical industry as a whole shows no immediate signs of reversing the decline in productivity observed over the last several decades. As an example, if the field of cheminformatics had been able to reduce the number of compounds needed to produce a drug by even 10−20% (presumably by accurately predicting the pharmacological properties of drug candidates), then we should be observing decreased costs and cycle times across the pharmaceutical industry. The fact is that, to date, there is little to no objective evidence to suggest that systematic application of cheminformatic principles (or any other relatively mature technology for that matter) has *increased* overall pharmaceutical productivity. An interesting recent analysis does suggest that, at least as a whole, the industry has become more efficient since the increase in the cost of drug discovery has outpaced the decline in output,[8] but this is an unsatisfying argument for the impact of any individual technology. There are many potential explanations for the lack of unambiguous, significant, and global impact of cheminformatic research on Discovery productivity. One possibility is that, given the long lag time between the Discovery cycle and the market, there simply has not been sufficient time to measure the impact of cheminformatics on the development of new chemical entities (NCEs). An argument has been made that this may very well be the case for other technologies, such as high-throughput screening,[9−11] and perhaps cheminformatics will have its day. Another explanation may be that the existing tools in fact perform their predictive function quite well but are not utilized in such a way that their full impact can be realized. This would speak to organizational or cultural issues that limit effectively capitalizing on the power of these approaches. Yet another possibility is that the existing cheminformatic tools or approaches are simply not accurate enough (in aggregate) to make much difference in such a highly complex and risky endeavor as new drug discovery.

Interestingly, we believe that there is evidence for all three explanations, which may result from scientific, cultural, or pragmatic issues, and the remainder of the manuscript will explore these in more detail. As a conceptual aid, the manuscript is divided into four sections that attempt to divide the universe of cheminformatic tools not only by our ability to

*To whom correspondence should be addressed. Phone: (847) 937-0368. Fax: (847) 938-2478. E-mail: philip.hajduk@abbott.com.

 Published on Web 03/24/2010

Table 1. Cheminformatic Tools for Abbott Medicinal Chemists

tool	utility	deployment	key references
Pipeline Pilot	used for calculating the vast majority of physicochemical properties and ligand efficiencies	web services	Accelrys http://accelrys.com/
Property Calculation page	calculates a standard subset of useful physicochemical properties, along with substructure search and clustering capabilities	web tool	
LeadHopper	combines 2D (ECFP6) and 3D (ROCS) methods for compound similarity searches	web tool	Muchmore (2008)[68]
RocsOverlay	provides 3D overlays of multiple input query molecules using the program ROCS	web tool	Grant (2007)[84]
DrugGuru	based on a query structure, generates a list of potential bioisosteric replacements	web tool	Stewart (2006)[54]
RocsDock	enables the generation of models of compounds docked to their receptor driven first by 3D ligand overlap and then simple minimization in the active site	web tool	Nichols, et al (2009)[75]
PyMol	general purpose molecular visualization. Customizable platform for application support and interface	desktop application/web client	DeLano (2002) http://www.pymol.org

actually engage in meaningful calculations but also our ability to interpret and apply the results. We have structured this according to a rubric used in military and legal circles and widely popularized by the United States Secretary of Defense Donald Rumsfeld in 2002: The *Known Knowns*, the *Known Unknowns*, the *Unknown Knowns*, and the *Unknown Unknowns*. This structure compels us to not only describe what we think we (or others) know, but also (and perhaps more importantly) what we do *not* know. It also forces us to ask what it is that is not known: Is it the method of calculation itself? Is it the underlying data quantity or quality? Is it the complexity of the system that defies simple analyses? For the most part, we have limited our discussion to tools or applications that can be used throughout the medicinal chemistry community (a partial listing of tools available to Abbott chemists is given in Table 1), with less attention being given to approaches that require expert knowledge in computational modeling (such as quantum mechanical calculations or complex virtual screening exercises) and that are best left to the experts. We hope that straightforward and honest (if incomplete) discussions of some of these questions will be of tremendous value to the medicinal chemist trying to use these tools in drug discovery programs.

Known Knowns

The *Known Knowns* are things that are (or should be) part of virtually every cheminformatic analysis performed on relatively large sets of molecules. These are calculations or analyses that we "know" how to perform and "know" how to interpret. As will become clear below, very few things can be confidently placed in this category. In fact, of the whole array of cheminformatic possibilities, we can only place three items in the *Known Knowns*: molecular weight, ligand binding efficiency, and substructure searching.

Molecular Weight and Atom Counts. The size of a molecule is straightforward to calculate, using either the molecular mass or the number of atoms. The utility of monitoring molecular weight (MW) and the counts of certain atoms (e.g., nitrogen and oxygen counts) during lead selection and lead optimization increased significantly after Lipinski's landmark publication correlating increased MW and atom counts with increased risk of clinical failure,[1] primarily due to low oral absorption. Subsequent studies have validated this observation[12] or have found that the number of rotatable bonds can serve as a surrogate marker for size.[13] Molecular weight has since become enshrined as one of the "rules of 5" (see Table 2), predicting an increased risk of clinical failure for compounds with molecular weights in excess of 500 Da.[12] While a number of studies have cautioned against overstrict application of these principles,[14,15] there is general acceptance in the scientific community that larger molecules will have reduced clinical success rates, and accordingly, one way to increase clinical success rates is to focus on making smaller compounds.

This general acceptance of the link between MW and oral absorption makes it one of the clearest examples of how difficult it is to rigorously apply cheminformatic principles in a pharmaceutical setting. For example, some protein targets (such as protein−protein interactions and peptidergic GPCRs) are simply not amenable to being targeted with "rule of 5" compliant compounds, and instead, general guidelines for the expected molecular properties of compounds targeting certain protein families have been described.[16] Thus, to address less tractable targets, larger, more lipophilic compounds are required, and one will need to find other ways to increase oral absorption (and overall clinical success) other than decreasing molecular size.[14] There are also the notable "exceptions" to the "rule of five", such as natural products, which exhibit good drug properties in spite of what might be predicted.[17] In fact, natural products were excluded from Lipinski's initial analysis for this very reason, suggesting that properties other than simple molecular weight are more important for achieving good oral bioavailability.[18] So, a "hard ceiling" of 500 Da is inappropriate in some settings while perhaps appropriate

Table 2. Cheminformatic Rules-Of-Thumb for Hit Selection and Lead Optimization

parameter	rules-of-thumb	comment	programs	key references
oral bioavailability ("rule of 5")	MW ≤ 500 Da ClogP ≤ 5 H-bond donors ≤ 5 #(N + O) ≤ 10	violation of these limits decreases oral bioavailability	Biobyte ClogP[85,86] or ACD LogP v4.0[12]	Lipinski (1997)[1] Wenlock (2003)[12]
oral bioavailability	Nrot ≤10 PSA ≤ 140 Å^2	violation of these limits decreases oral bioavailability	tPSA[62] (nitrogen and oxygen only)	Veber (2002)[13]
oral bioavailability ("Golden Triangle")	MW ≤ 500 variable LogD (LogD range: 0 − 5)	violation of these limits decreases oral bioavailability	experimental LogD	Johnson (2009)[35]
toxicity	ClogP ≤ 3 PSA ≥ 75 Å^2	violation of these limits increases the risk of toxicity	Biobyte ClogP v4.3[85] tPSA[62] (nitrogen and oxygen only)	Hughes (2008)[2]
toxicity	LLE ≥ 5	low ligand-lipophilicity efficiency can lead to increased promiscuity	Biobyte ClogP[85]	Leeson (2007)[19] Leach (2006)[23]
membrane permeability	PSA ≤ 120 Å^2	violation of this limit decreases membrane permeability	Quanta 3D (nitrogen and oxygen only)	Kelder (1999)[61]
membrane permeability	MW ≤ 500 variable LogD (LogD range: 0.5 − 5)	violation of these limits decreases membrane permeability	ACD PhysChem Batch[87] or AZlogD[88]	Bhal (2007)[34] Waring (2009)[36]
blood−brain barrier penetration	PSA ≤ 70 Å^2	violation of this limit decreases brain penetration	Quanta 3D (nitrogen and oxygen only)	Kelder (1999)[61]
solubility	Fsp3 ≥ 0.4	increased fraction of sp3 hybridized carbons (Fsp3) increases solubility	Pipeline Pilot 7.5	Lovering (2009)[51]
general "developability"	number of aromatic rings ≤ 3	increase in aromatic ring count decreases solubility and increases protein binding	none listed	Ritchie (2009)[52]

in others. This has led to a nonsystematic and perhaps even haphazard application of this rule in many Discovery settings (even within the same group of medicinal chemists), making it difficult to assess its impact on productivity. Support for this view comes from a recent analysis from AstraZeneca, where the physicochemical properties of patented compounds from four major pharmaceutical companies were compared.[19] It was concluded that a large fraction of compounds emerging from these patents violate simple drug-like property filters and therefore carry increased clinical risk, despite the ready availability and acceptance of property filters in compound design. Importantly, these trends were even observed when different companies produced compounds against the same target, suggesting that the nonsystematic use of MW (or other property) filters was more cultural (i.e., some companies paid more attention than others) than scientific (i.e., the target required violating the "rule of 5").

The fact that chemists at different companies vary significantly in their adherence to certain "rules" or "guidelines" is an incredibly important statement about the use of cheminformatic approaches in drug discovery and has to do with the actual versus perceived risk of taking one path over another and how cheminformatic approaches can modulate that risk. An illustrative example is the use of cheminformatic filters to triage (i.e., remove compounds from) high-throughput screening (HTS) hit lists. Identifying the truly promising dozen or so hit series from a typical HTS hit list of thousands of compounds requires deprioritizing ~90% of the actives, which can involve activity confirmation, structure confirmation, selectivity panels, limited HT-ADME data, and more. Suppose that a simple molecular weight filter of 500 Da removed ~50% of the initial hits. Should these compounds be removed from further consideration and therefore reduce the hit-to-lead burden by a factor of 2? Consistent with external reports,[20] we have found that high molecular weight compounds are more likely to be false positives than low molecular weight compounds, and therefore, engaging in hit-to-lead activities on compounds more likely to be artifacts misdirects precious resources. However, a common contrarian response is that perhaps there are interesting actives in this 50% that can be appropriately "down-sized" during lead optimization. This is most certainly true in some cases. However, it is also true (in this hypothetical scenario) that 50% of the actives *already pass* this property filter and may represent even higher quality leads (potentially not requiring the degree of optimization as the larger molecules). So, removing the high MW molecules is low risk so long as a sufficient quantity and quality of *bona fide* leads can be found in the other half of the hit set. Thus, application of cheminformatic filters can make the overall process of hit triage more efficient (which is good) with an increased risk of missing some number of potentially interesting hits (which could be bad).

Ligand Efficiency. The concept of ligand efficiency has emerged as one approach to meaningfully interpret molecular size by balancing the size of the molecule against its potency.[21] Thus, a larger molecule may be viewed favorably if its potency is subnanomolar but unfavorably if it only exhibits micromolar activity. One of the first equations for calculating ligand efficiency was given by Hopkins,[22] where the binding energy ΔG was divided by the number of heavy atoms. Subsequent formulations have involved using the pK_I (the negative base-10 logarithm of the K_I or IC_{50}) divided by either the number of heavy atoms[23] or the molecular weight.[24] The latter formulations are preferred, as the pK_I (or pIC_{50}) is readily calculated and interpreted by the medicinal chemist, while the free energy of binding (ΔG) is not commonly used. The practice of using ligand efficiency to prioritize compounds puts all molecules "on the same playing

field," so to speak, such that highly efficient fragment hits can be rationally compared with larger (and proportionately more potent) molecules. This is especially important when both conventional HTS and fragment-based lead discovery strategies are employed against the same target. As with molecular properties, the maximal binding efficiency that can be achieved with any small molecule will depend on the target type, but useful ranges can be defined using a desired affinity and proteomic-based molecular properties.[16] We have adopted this approach for hit triage from HTS, where compounds that fall within the optimal expected ranges of ligand efficiency for that particular target class[16] are prioritized for further evaluation. While we have placed ligand efficiency in the *Known Known* category (as the mass is known exactly and we can experimentally determine potency), we must recognize that the biological activity is in fact not fully "known" in the sense that assays are subject to error and (more significantly) do not always reflect the relevant biology. Nonetheless, we usually have enough confidence in the experimental measurements to apply these principles as *Known Knowns*.

Ligand efficiency has its greatest utility at the stage of hit or lead selection, where only the most efficient (and potentially most "optimizable") compounds are taken forward. However, it is possible to use the concept of efficiency throughout optimization by evaluating the impact that any specific substituent makes on both potency and size. For example, achieving a 5-fold gain in potency by adding 120 Da to your molecule may be less than ideal.[25] This analysis has been coined "group efficiency," where the efficiency of binding of any single substituent can be analyzed.[26]

Substructure Searching. Substructure searching (SSS) is another foundational cheminformatic tool in the practice of medicinal chemistry. SSS usually attempts to address one of two questions: (1) what other molecules contain a substructure of interest, and (2) what molecules do not contain any of these substructures? The first is typically used for finding analogs of known actives for the development of structure−activity relationships (SAR), while the latter is typically used for flagging compounds that contain certain problematic substructures. This tool is placed in the *Known Known* category because we generally know how to perform substructure searches (most chemical databases are SSS-enabled) and understand what we get (i.e., this molecule contains a pyridine). However, while substructure searching for SAR development is effective, it can be time-consuming and subjective. As an example, consider the molecule in Figure 1. Dozens of different types of substructure searches can be done on this single molecule by varying which rings and exocyclic substituents are retained and whether multiple atom types are allowed at specific positions. It is usually up to the medicinal chemist and his or her knowledge of the existing SAR and chemistry to define the substructures and perform these searches. This is especially true at the stage of hit triage, where it is not at all clear what substituents or atoms are required for activity (part of the "pharmacophore") and which portions of the molecule can be modified. As a result, substructure searching in the quest for new active molecules is as much of an art as it is a science.

In addition to searching for specific substructures to fill out SAR, most pharmaceutical companies have a list of substructures that are undesired, as they carry some liability that may be difficult, if not impossible, to overcome. Such liabilities can be screening artifacts,[27] compound

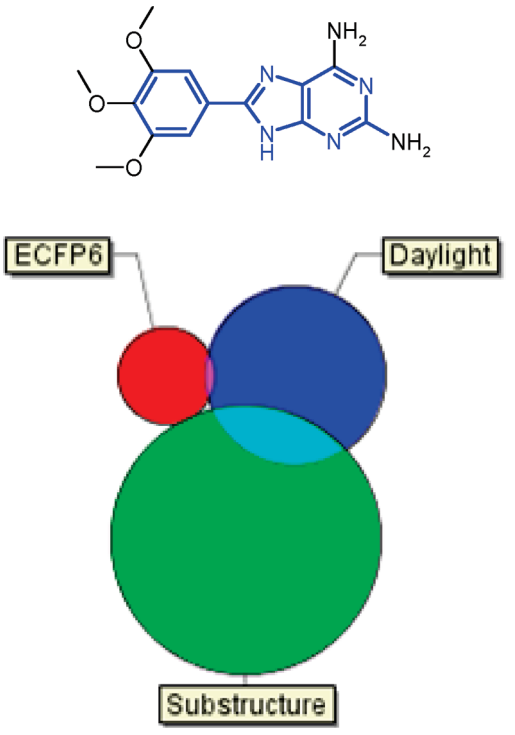

Figure 1. Structure of a hypothetical HTS hit. A substructure search around the structure shown in blue against the Abbott corporate collection yielded 209 unique hits (green), while similarity searches using Daylight fingerprints (blue) or ECFP6 fingerprints (red) yielded 94 and 27 hits, respectively.

reactivity,[28−30] and toxicity,[31,32] among others. It is interesting to note that there is actually a great deal of consensus that certain functionalities are to be avoided in HTS campaigns,[33] to the extent that Scitegic even supplies an "HTS Filter" component as part of its standard array of property filters. Thus, large lists of compounds can be passed through cheminformatic protocols to determine whether the candidate molecules contain any functionalities of concern. However, as with molecular weight, what action is taken on the basis of this information varies from the liberal (e.g., simply flagging the molecule for a chemists' visual inspection) to the conservative (e.g., removing, physically or otherwise, the compounds from further consideration). At Abbott, like many other pharmaceutical companies, we employ a balance of both approaches, where compounds containing certain substructures have been physically removed from the screening decks, while other substructures are simply flagged to make the medicinal chemist aware of an issue that may need to be addressed during lead optimization.

Unknown Knowns

Unknown Knowns are things "that we don't know or intentionally refuse to acknowledge that we know" (Wikipedia). In the realm of cheminformatics, these would be things that we would claim we do not yet know how to do but which would be very useful. In many cases, models or other computational tools may be available but most of these are "known" to be less than useful. Perhaps controversially, this category also includes properties or attributes for which we are perhaps overly confident in trying to predict, in a sense, failing to acknowledge the limits of our tools and methods.

Calculated LogP and LogD. The partition coefficient, LogP and the related distribution coefficient, LogD are important pharmacokinetic values for medicinal chemists since they serve as estimates for the distribution of drug substances in the human body. LogP is the log of the measured ratio of compound in a nonpolar organic solvent (such as octanol) to the concentration of the compound in a nonionized aqueous solution. This ratio is a useful measure of the relative hydrophobicity/hydrophilicity of a compound, and is an important property in predicting ADME characteristics of a compound (see Table 2). Given the importance of this coefficient, a number of approaches to its calculation have been pursued, and high calculated values for LogP have been shown to correlate with decreased oral bioavailability (as part of the "rule of five"[1]), increased promiscuity,[20] and increased risk of toxicity.[2] It has recently been demonstrated that LogD is perhaps a better predictor for drug-likeness than LogP,[34−36] with modified rules-of-thumb for permeability (see Table 2), although LogP is likely still more important for overall toxicity. As with molecular weight and the concept of ligand efficiency, the lipophilicity of a compound can be balanced by its absolute potency using a term alternatively called lipid efficiency (LipE) or ligand−lipophilicity efficiency (LLE), which simply subtracts the calculated LogP (or logD) from the pIC_{50}.[19,37]

As a result of these and other studies, the calculated LogP parameter is routinely included in cheminformatic analyses throughout the pharmaceutical industry. However, it may be surprising to many that even with the wealth of experimental LogP data that have been acquired over the years, the existing models for calculating LogP are modest at best. In fact, a recent survey of 30 different methods for predicting LogP suggests that even the very best performing models achieve a mean error of 1 log unit on naive data sets.[38] This certainly cautions against hard limits on calculated LogP for cheminformatic filtering, as many molecules will be incorrectly classified. The review also highlights significant differences between the different programs for calculating LogP, such that applying recommendations from one analysis determined using one LogP calculation method may be invalid (or seriously misleading) if a different method for calculating LogP is used. This can be appreciated from Figure 2A, where two different methods for calculating LogP on the same set of 10,000 compounds are compared. While highly correlated (R^2 = 0.79), the mean absolute error (MAE) between these two methods is 0.66 log units, with 22% of the compounds exhibiting more than a log unit difference and 3% of the compounds differing by more than 2 log units. The situation is further complicated when tautomers are considered, as different tautomers of the same molecule can have significantly different calculated molecular properties.[39] While we strongly recommend that the cautions regarding high LogP (or LogD) be incorporated into hit selection and lead optimization, cheminformaticians, and medicinal chemists alike can all too often fall into the trap of calculating what is easiest (e.g., I have this program but not that program) as opposed to what is appropriate (e.g., I have to find a way to use Method X). The ready availability of desktop, enterprise-wide cheminformatic toolkits (e.g., Pipeline Pilot from Scitegic) makes it even more important that the end user understand what is actually being calculated before implementing any filtering protocols. Of course, the required level of accuracy will be different when one is simply assessing trends in hydrophobicity (for which the current methods are more than suitable) versus using calculated logP as a "hard" cutoff for filtering (where many compounds may be misclassified). An alternative approach is to actually use experimental LogD values for a few members of a series to either validate the applicability of the general model to the series of interest or to develop local models that may exhibit superior performance, but this is typically outside the domain of most medicinal chemists. It is for these reasons that calculated LogP is, in our opinion, squarely in the *Unknown Known* category, we can often predict (or apply) overconfidently and not pay adequate attention to the well-established errors associated with each technique.

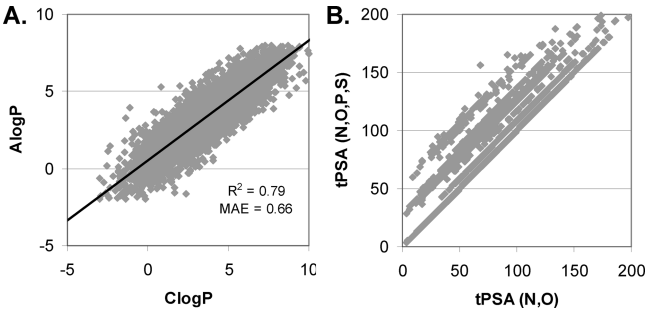

Figure 2. Comparison of (A) calculated logP and (B) polar surface area using different methods as described in the text. ClogP refers to the Biobyte ClogP,[85] while AlogP is as implemented in Pipeline Pilot v7.5. tPSA was calculated in Pipeline Pilot v7.5 using the "Surface Area and Volume" component. The default setting for this component utilizes nitrogen, oxygen, sulfur, and phosphorus atoms in the tPSA calculation.

Solubility. High solubility of test compounds in intestinal fluids provides a useful concentration gradient that aids the absorption of orally administered compound. Furthermore, enhanced solubility of the compounds in typical in vitro assay vehicles (DMSO/aqueous buffer) precludes the false negatives or even false positives resulting from compound aggregation and precipitation. After all, aggregation/precipitation will lead to lower effective concentrations of the test compound or may cause the protein to coprecipitate.[27] The search for potent compounds often involves the incorporation of lipophilic pharmacophores with the resulting compounds displaying poor aqueous solubility. The promise of physiochemical property predictors to allow the modification of lead structures to meet specific solubility targets would facilitate the design of compounds to achieve appropriate exposure at a reasonable dose and generally enhance the quality of in vitro data sets.

A survey of the available cheminformatics tools used to predict solubility reveals a strong dependence on the general solubility equation:[40,41]

$$\text{Log } S_w = -0.01 \times (\text{MP} - 25) - \text{ClogP} + 0.5$$

where S_w is the aqueous solubility, MP is the melting temperature (in degrees Celsius), and ClogP is the octanol−water partition coefficient. An obvious initial complication for calculating solubility is that different polymorphs of the same compound will exhibit different melting temperatures. However, in practice, unless an abnormal melting temperature is suspected, most of the useful enhancement of solubility comes from reducing the logP. Given the errors inherent in predicting a LogP value as described above, it is not

unusual for a medicinal chemist to be rather dismissive of the predictive capabilities of solubility models, especially since charged molecules need an additional correction for pK_a (where the errors in calculated logP and pK_a are then compounded).[41] As one author points out, most of the solubility estimations only work reliably on noncharged compounds.[42] This is particularly troublesome when you consider that over 80% of marketed drugs contain ionizable fragments. Many medicinal chemists will likely be intrigued by the results of an exercise characterized as a "Solubility Challenge".[43] In this contest, researchers were challenged to predict the solubilities of 32 compounds given a database of 100 intrinsic solubilities measured for drug like molecules. Not surprisingly, one analysis of the challenge[44] concluded that most methods yield generally poor results (with mean errors in excess of 1 log unit), with widely varying behavior based on the type of compound being predicted (e.g., charged vs noncharged). Another interesting conclusion from this work was that a more sophisticated model did not necessarily perform better than simpler models. The authors therefore concluded that "the limitations in the current ability to predict aqueous solubility are probably not a result of inadequate modeling methodologies, but are more probably a result of an insufficient appreciation for the complexity of the solubility phenomenon".[44] A similar conclusion was reached in the comprehensive study of LogP predictions described above.[38] Thus, for many parameters that we would desire to predict accurately, there are certain fundamental limits of our understanding that preclude greater accuracy, even if more data could be collected to power model development. We know this, but in our zeal to calculate and analyze we often forget.

Despite these limitations, predicting solubility can be useful even in the context of large errors. For example, trends in solubility can be assessed, such that absolute solubility is not required, but whether one compound in a series is more soluble than another. This is especially valid since the absolute solubility in water will most certainly be different (and usually lower) than in vivo, where albumin and other serum components can aid the solubilization of many hydrophobic compounds. In addition to calculating absolute solubility, several recent analyses have looked at the effects of single changes to a compound and their relative impact on potency, solubility, and other properties.[45−50] By building up a sufficiently large database of these "matched molecular pairs", certain "rules-of-thumb" can be derived from these data analyses to identify transformations that consistently bias toward or away from increased solubility. Such analyses also reveal other general trends for increasing solubility, such as increasing the number of hydrogen bond donor and acceptor moieties, increasing the number of rotatable bonds, reducing symmetry, and increasing saturation.[51]

Plasma Protein Binding. The ability of a compound to bind to plasma proteins can significantly modulate its in vivo activity. In some cases, this can be beneficial as plasma proteins can serve as useful vehicles to transport lipophilic compounds through the circulatory system to sites of biological activity. Furthermore, proteins such as alpha-1-acid glycoprotein and human serum albumin (HSA) can reduce the clearance of compounds, thereby prolonging the pharmacokinetic action of the compound. However, high affinity binding to serum proteins can reduce the level of unbound (and, therefore, pharmacologically active) compound available to its target. Thus, it is often the case that the medicinal chemist's job is to find the appropriate balance between low and high affinity binding. Unfortunately, predicting the affinity of a compound for albumin is notoriously difficult. The general rule is that lipophilic, acidic compounds will bind to albumin, and affinity will correlate with hydrophobicity. This is confirmed by a recent report correlating protein binding with the number of aromatic rings in a compound, which is another predictor of successful drug development.[52] Therefore, incorporating anionic amines or reducing logP are avenues for reducing the level of protein binding.[53] Standard approaches for this have been incorporated into our bioisostere tool DrugGuru.[54] While there have been a few reports of rational, structure-based approaches to reducing binding to albumin,[55,56] high protein flexibility, and multiple ligand-binding sites make it a formidable challenge both for obtaining high resolution structures and for executing structure-based drug design.

In Vivo ADME. Ultimately during the lead optimization process, the medicinal chemist is faced with the obstacle of ensuring appropriate systemic exposure of their candidate compounds in addition to affinity for the molecular target. While many tools exist to predict interactions with the protein target of interest, there are considerably fewer tools for predicting the pharmacokinetics of the newly designed compounds. Structure−activity relationship papers produced by medicinal chemists reveal a systematic optimization of the candidate compound for the protein of interest, which is often translated incorrectly to suggest that the design process involves only optimization of activity for the target protein. In fact, the oral exposure of a compound is influenced by a number of important factors, such as intestinal fluid solubility, epithelial permeability, metabolic clearance, transporters, renal clearance, and others. Ironically, the optimization of compounds for ADME properties often results in more hydrophilic, less membrane permeable compounds, negatively affecting oral exposure and affinity for the target of interest. Therefore, the proper prediction of these ADME properties could dramatically enhance the development of safe, effective, and orally available drugs. As databases containing these experimental measures increase (both in the public domain and proprietary databases), numerous predictive models have been described (for good overviews, see refs 57−59). However, the complexity of these biological end points has hampered the development of accurate, reliable, global models, and much research is still needed.[60] As discussed above for solubility, more success has been realized in the development of local or fragment-based models, where the average effects of specific substituents can be studied in isolation. Such analyses have been reported for substituent effects on metabolic stability,[50] Cyp450 and hERG inhibition,[49] and protein binding.[46,47] We anxiously await continued progress in this field.

Known Unknowns

Known Unknowns refer to "circumstances or outcomes that are known to be possible, but it is unknown whether or not they will be realized" (Wikipedia). From a cheminformatics perspective, these are things that we "know" how to calculate (or at least generate a number), but their usefulness must be qualified (i.e., unknown whether usefulness will be realized). This section contains by far the largest number of cheminformatic approaches.

Polar Surface Area. The polar surface area (PSA) of a molecule is another molecular descriptor that is ubiquitously

Table 3. Partial List of Chemical Similarity Programs

program name	type of similarity measure	source
Pipeline Pilot (ECFP(2−12), FCFP(2−12), etc)	connectivity fingerprint	www.accelrys.com
MACCS Keys	2D fingerprint	www.symyx.com
Lingos	2D fingerprint	www.eyesopen.com
MCS	maximum common substructure	www.chemaxon.com
Unity 2D/merq	2D fingerprint	www.tripos.com, www.daylight.com
Phase	3D shape	www.schrodinger.com
ROCS	3D shape and chemical group similarity	www.eyesopen.com
EON	3D electrostatic similarity	www.eyesopen.com
USR	3D Shape similarity	http://www.isis-innovation.com/licensing/2932.html

used in drug discovery. PSA has been shown to correlate well with the passive transport of compounds across biological membranes and therefore has been used as an indicator of the propensity of a compounds to penetrate the blood−brain barrier.[61] Rules-of-thumb from these studies suggest that orally bioavailable compounds have PSA values less than 120 Å^2, while compounds that penetrate the blood−brain barrier typically have PSA values less than 70 Å^2 (see Table 2). PSA has also been implicated in modulating general toxicity, in combination with ClogP,[2] where compounds with PSA values less than 75 Å^2 and ClogP values greater than 3 are significantly more likely to be toxic.

The classical calculation of the surface area is done by first calculating a 3D structure of the compound, and then using a molecular surface representation to quantify the surface area surface associated with polar atoms. In the original paper describing this work, it was noted that the surface area calculated in this fashion did not show a high degree of dependence upon the conformation of the 3D representation of the compound. Subsequently, Ertl et al. proposed a simplified version of the calculation, topological surface area (TPSA), which does not require the calculation of the molecular surface area to derive the PSA[62] and, therefore, is faster than the original calculation. The two methods have been shown to correlate well with each other (with mean absolute errors less than 6 Å^2), and therefore, the TPSA calculation is often substituted for the 3D methods. Because of the high degree of concordance between the two methods, we have recently replaced all internal calculations of PSA with TPSA. Surprisingly, this was a source of intense discussion within our medicinal chemistry community. First, the new values did not correspond exactly to previous studies performed with 3D PSA, which meant that some ongoing studies needed to be updated with new values. However, there was also engaged debate around which calculation was "correct." Somewhat disconcertingly, we had to confess that there is no "correct" answer, in that polar surface area is a complete theoretical abstraction with no experimental "truth" against which to compare (as opposed to logP and solubility, for example). The greatest argument for its validity is the strong empirical correlations with observable parameters. It is for this reason that we placed PSA in the *Known Unknown* category, as we can calculate, with infinite precision, a parameter that has no experimental correlate, but that nonetheless has significant empirical value.

One final note of caution when using PSA is the classification of sulfur and phosphorus atoms as either "polar" or "nonpolar".[62] Again, this is a matter of subjectivity with no "correct" answer, but one can obtain quite different results. Shown in Figure 2B are TPSA values for 10000 compounds treating either the set {N, O} or the set {N, O, S, P} as polar. Systematic deviations as large as 60 Å^2 can be observed for compounds containing sulfur or phosphorus atoms in certain contexts. As sulfur is present in 25−30% of common drug-like molecules, care must be taken to make sure that the settings on programs correspond to what is actually desired.

Chemical Similarity. Chemical similarity is an important concept, particularly in drug discovery. Structure−activity relationships (SAR) are typically derived for sets of similar molecules, and chemical similarity is the basis for clustering molecules into structurally related groups. The guiding principle for chemical similarity is that structurally similar molecules should exhibit similar biological activities.[63] There are numerous ways of calculating chemical similarity, but similarity programs can be categorized in two general groups: those that use 2D information and those that use 3D information. A listing of software typically used for these calculations is given in Table 3.

Deciding whether two molecules are similar is much like trying to decide whether something is beautiful. There are no concrete definitions, and most chemists take an "I know it when I see it" attitude (attributed to United States Justice Potter Stewart, concurring opinion in *Jacobellis v. Ohio* 378 U.S. 184 (1964), regarding possible obscenity in The Lovers). From a cheminformatics perspective, then, chemical similarity is a *Known Unknown* because we know we do not know it. Unfortunately, this is not terribly helpful for the cheminformatician trying to wade through hundreds of thousands of compounds to identify similar or dissimilar molecules. To address large data sets, the vast majority of chemical similarity programs take a structure and break it into "bits" of information (see Figure 3). For example, compounds with an amide may turn the amide "bit" to "on," while compounds lacking an amide leave this "bit" off. Once a molecule is defined by a string of bits, then some very straightforward analytical expressions can be used to determine how related the two compounds are, with the most famous being the Tanimoto coefficient (see Figure 3). The problem (and thus the large number of chemical similarity approaches) is how you define the "bits" of information that comprise the molecular "gene". These can yield wildly different results, which can be different yet again from substructure searching. An example of this is given in Figure 1, where similar structures for a single molecule were searched for in our corporate database using three methods. Manual searching for the substructure shown in blue yielded 209 hits, while similarity searches using either Daylight[64] or extended connectivity fingerprints (ECFP6)[65] yielded 94 and 27 molecules, respectively. The Venn diagram in Figure 1 shows that, for this particular molecule, there are very few retrieved molecules in common between any two of these three search methods! This does not at all mean that the

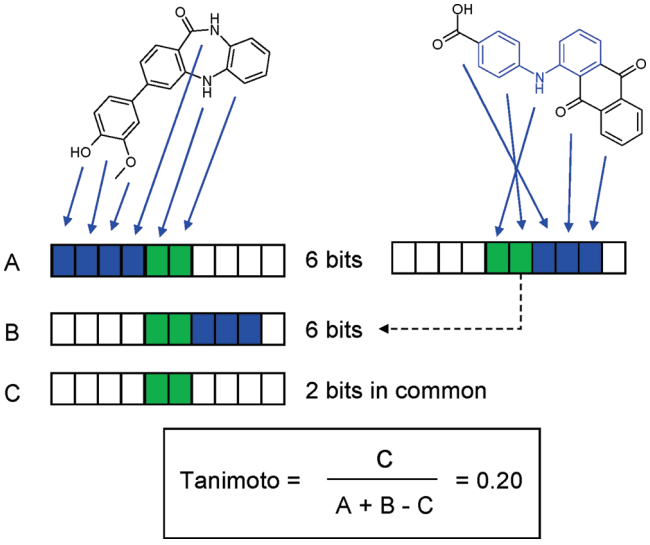

Figure 3. Schematic illustration of the calculation of the Tanimoto coefficient.

searches failed, but simply that each method targeted different aspects of the information contained in the query molecule. While this example was one of the most extreme cases we could find, it is almost universally true that each search method will return a slightly different list of compounds.

This example poses two challenges for the medicinal chemist trying to utilize chemical similarity programs. First, what level of Tanimoto coefficient is appropriate to define "similar," and, second, how does one combine the results of multiple similarity search algorithms to provide a more comprehensive list of similar structures from a single search? Several approaches are available for combining different similarity search results, including consensus scoring, rank normalization, and others.[66,67] We have described a probabilistic framework that reduces the resulting Tanimoto coefficients to probabilities that a compound will be active, and that can be used to combine different metrics by employing Belief Theory.[68] Significantly, this approach (which we call "LeadHopper") is available to the entire medicinal chemistry community through a web-based protocol that returns a single list of compounds (with their likelihood of being active) from multiple search queries. An important aspect of this probabilistic framework is that molecules which fall below a defined probability of being active are not returned (such that it is possible to get no retrieved structures if the query is a true structural singleton) and *all* molecules are returned above this threshold. This is very different than simply returning the top 1000 compounds, regardless of absolute level of similarity, and gives an initial read as to how densely populated the chemical space is around the query molecule.

Three-Dimensional Compound Overlays. It has long been appreciated that the 2D similarity algorithms can miss compounds that have significant differences in their atom and bond topology, but have similar overall shape and electrostatic properties. The 3D similarity programs, such as ROCS,[69] SURFLEX,[70] and others, exploit molecular shape similarity, in combination with chemical pattern matching to score compounds relative to one another in terms of 3D similarity. To generate a similarity score, these algorithms must first identify the best achievable shape and electrostatic overlay of the compounds in three dimensions, complete with conformational searches around rotatable bonds. Thus, these approaches allow visualization of the best matched poses, which can provide invaluable (and often surprising) insights into how two different molecules may bind to the same receptor site. We have therefore provided a simple web tool (which we call "RocsOverlay") to allow medicinal chemists to input any two structures and obtain interactive 3D overlays. It must be stated that tools of this type will return a 3D overlay in every case, even if it does not ultimately make sense, so care must be taken in interpreting the results (in our internal deployment of these results, data to aid this interpretation is supplied). It must also be mentioned that there is no claim that the returned conformations are the *bioactive* conformations, but simply the conformations that allow the best overlay between the two molecules. So, we can calculate a "best" overlay, but we cannot really "know" if we are right. Thankfully, it has been our experience that you do not have to be "right" all of the time, you simply have to be "useful" most of the time.

Bioisosteres. What naturally follows from visualizing the superposition of two molecules is "what can replace what?" or bioisosterism. Bioisosterism is highly related to chemical similarity, with the distinction that bioisostere replacement typically involves replacing only one part or functional group of a larger molecule, whereas chemical similarity generally deals with the entire molecule. Thus, bioisosteric compounds are those related to each other by the exchange of atoms or groups of atoms that are similar, or known to have similar chemical properties (e.g., size, shape, electrostatics, etc.), but with potentially improved biological properties.[71] The identification of bioisosteric pairs is a cherished tradition within medicinal chemistry communities. Many isosteres are commonly known (e.g., a tetrazole replacing a carboxylic acid), while others are less obvious and are often only identified empirically throughout the course of lead optimization. To capture this colloquial knowledge and make it available to our entire medicinal chemistry community, we have cataloged both bioisostere and other common transformational "rules" that have been discovered either externally or internally and have developed a tool (which we call "DrugGuru") that takes an input molecule and changes the structure based on an ever-growing list of transformations (which currently stands at more than 250 rules that have been successfully exemplified).[54]

When searching for bioisosteres, one is typically looking to maintain some quality of the compound (e.g., binding affinity to the target of interest), while changing other properties (e.g., decreasing LogP). It is very straightforward to assess the impact of structural transformation on any number of calculated physicochemical properties, as described above. However, combining the ability to generate transformed molecules and overlay them back onto the parent structure in three dimensions (see Figure 4) is a powerful approach for evaluating the likelihood that the proposed bioisostere will retain biological activity. Of course, likelihoods only reduce to reality when compounds are actually synthesized, and it is the task of the medicinal chemist to use these tools to focus his or her attention on the most likely candidates. The challenge of predicting the synthesizability of the molecules proposed by such software is yet another *Known Unknown*, and the reader is referred to a number of other reviews on that topic.[72,73]

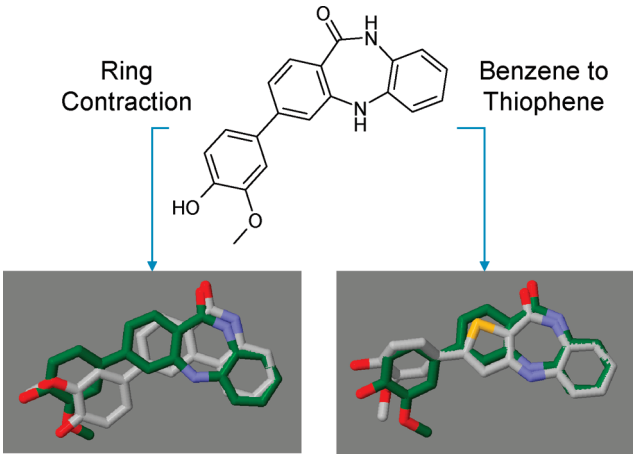

Figure 4. Example of bioisostere replacements and subsequent 3D overlays for visual inspection. In this case, the "ring contraction" rule significantly distorts the 3D geometry of the ligand, while the "benzene to thiophene" rule maintains a good overlay.

Binding Pose Prediction. During a structure-based drug design (SBDD) effort, it often is desirable to generate the bioactive conformation of a small molecule ligand relative to its target before synthesis is initiated to understand whether the structural change is compatible with the protein binding site. Depending upon the level of information available, the generation of accurate binding poses can be simple or very daunting. In the most difficult cases, an expert modeler uses his or her chemical intuition and manually generates likely interaction poses, usually in combination with energy minimization routines to retain reasonable geometries. However, during lead optimization in a SBDD project it is likely that some structural data about the target and related bound ligands are available. In this case, there are programs that will allow the user to provide structural template information that allows the automated placement of compounds with some degree of confidence. The CORES approach, in which known fragments are used to initially place the compound into the active site of the drug target, has been used to generate accurate poses.[74] Recently, another approach has been described using the molecular shape of a known molecule, in combination with simple scoring functions, to guide the pose prediction.[75] This approach has been shown to be able to reproduce known poses with an average accuracy approaching 0.5 Å rms from the known binding poses.

Given the fact that available structural information on protein−ligand complexes can enable (mostly) reliable and automated docking of structurally related compounds, we have created a web-based tool called "RocsDock" and made it available to Abbott chemists. This tool allows the user to select an available protein−ligand complex (either experimental or modeled) and simply supply a structure for docking (in any variety of formats, such as SMILES, SD files, or compound registration codes). The initial pose generation is driven by shape and electrostatic overlay with the known bound ligand (using the program ROCS), followed by simple minimization in the protein. As with RocsOverlay, a docked pose is always returned. However, predefined thresholds for shape overlay with the known ligand and shape complementarity to the protein provide a readily interpretable level of confidence to the user. In line with the theme of this article, we make our best estimate of the correct pose using all available information, but it is up to the user to decide whether it is actually useful: a *Known Unknown*.

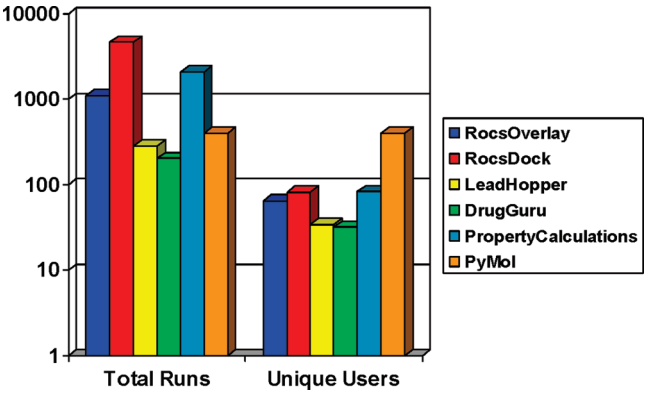

Figure 5. Usage statistics for Abbott cheminformatic tools in 2009.

Unknown Unknowns

Unknown Unknowns refer to "circumstances or outcomes that were not conceived of by an observer at a given point in time" (Wikipedia). Formally, this section should be blank, as it is impossible to describe that which we cannot conceive. Nonetheless, we propose that the *Unknown Unknowns* are "pipe-dreams" for the cheminformatician, the things that we have not yet figured out how to reliably roll out to the larger medicinal chemistry community and that we are not even sure should ever be rolled out (for a variety of reasons). One large category of computational work that we have left in this category is that of large-scale virtual screening and calculation of ligand-binding energies. There are certainly useful things that have and can be done with these methods, but their approachability and value as a useful tool for medicinal chemists is highly questionable, and we have chosen for the moment to keep this in the domain of the expert computational chemist. The same can be said for QSAR studies, where much useful work has been done (see www.qsar.org for the Cheminformatics and QSAR Society's homepage). The difficulty with enabling community-wide QSAR model development is not that they are too complex to develop (in fact, they are frightening easy to build), but that their reliability and domain of applicability (i.e., their usefulness against new chemical matter) is often exceedingly challenging to define.[76,77]

Another exciting area of research that one can envision engaging the larger community is that of systems biology and polypharmacology. Some of the recent work by Hopkins,[78] Shoichet,[79] and others provides an entirely new level of understanding SAR and bioactivity against entire ranges of protein targets, and providing tools to rationally approach systems biology will be a growing need for future pharmaceutical research. Building on this dream of being able to model systems biology is the simulation of drug effects on whole human tissues, where future cheminformatic models are actually representations of a human body. Modeling compound effects on organs and tissues is an active area of research,[80−83] but the technology is in its infancy and it is questionable whether it will reach the desktops of research medicinal chemists in the next decade. Nonetheless, we look forward to the day when the biophysical impact of an administered compound on every tissue and against every gene product will be reduced to a set of equations. Only then can the "admittedly absurd" proposition be realized, that a

research chemist can design a single compound and accurately predict whether or not it will be a drug.

Conclusions

There has been a significant increase in both the quantity of and access to cheminformatic tools to enable drug research. Unfortunately, many cheminformatic approaches simply overpromise and under-deliver and, therefore, do not improve productivity (and may even reduce it). As discussed, this is sometimes a scientific issue that can potentially be addressed by defining specific criteria for utilization. It is also the case that many well-established principles are either not properly disseminated to or are simply dismissed by many medicinal chemists. This is a cultural issue, both in terms of the available infrastructure for deploying cheminformatic tools and in the acceptance by the medicinal chemistry community of the underlying cheminformatic principles. Finally, it must be recognized that cheminformatics is a rapidly evolving field that requires vast amounts of information in order to construct robust models. The current explosion in chemical and biological data available in both public and private databases has driven some of the more recent and exciting developments in the field which may truly approach global applicability and utility. At Abbott, there is certainly no shortfall in the utilization of enterprise-wide cheminformatic tools made available by the Modeling and Cheminformatics groups, with hundreds of unique users and thousands of usages per year (see Figure 5). We fully expect these and other tools to tangibly impact Discovery productivity over the next several years.

Acknowledgment. The authors would like to acknowledge Drs. Donald Halbert, Stevan Djuric, Jonathan Greer, and Yvonne Martin for critical reading of the manuscript and Drs. David Price and Paul Leeson for providing details on their ClogP calculators.

Biographies

Steven W. Muchmore, Ph.D., is an Associate Research Fellow in Global Pharmaceutical Research and Development at Abbott Laboratories. Steve joined Abbott in 1994 as a Postdoctoral Research Fellow and in 1995 became a full-time scientist in Structural Biology. In 2001, he assumed leadership of Abbott's fledgling Computational Structural Biology group and in 2007 became the leader of Abbott's Cheminformatics group. His leadership has brought to Abbott new methods and technologies for doing computational ligand and structure-based drug design.

Jeremy J. Edmunds is the Director of Medicinal Chemistry for the Immunology therapeutic area at Abbott Laboratories. Here he is responsible for a group of medicinal and analytical chemists that design and synthesize compounds to treat diseases such as rheumatoid arthritis, multiple sclerosis, and asthma. He received his Ph.D. in organic chemistry at Imperial College, U.K. under the supervision of Professor William Motherwell. After postdoctoral studies with Professor Anthony Barrett, he joined Parke-Davis Pharmaceutical Research, division of Warner Lambert, in 1990. There he remained through the acquisition by Pfizer to lead a group of medicinal chemists as the Director of Cardiovascular Chemistry. He joined Abbott in 2006 and currently enjoys medicinal chemistry and preclinical studies at the Abbott Laboratories site in Worcester, MA.

Kent D. Stewart is Volwiler Research Fellow within the Structural Biology group in Abbott Global Pharmaceutical R&D. After receiving his Ph.D. in organic chemistry from the University of California, Los Angeles, under the direction of Donald J. Cram and postdoctoral work in biochemistry under Emil T. Kaiser, Jr. at Rockefeller University, he held several pharmaceutical and academic appointments before joining Abbott in 1992. He provides molecular modeling support in molecular design and protein engineering to discovery teams across all research divisions of Abbott.

Philip J. Hajduk is the Associate Director of Lead Discovery at Abbott Laboratories, where he is responsible for the High-Throughput Screening, Fragment-Screening, Cheminformatics, and Compound Management groups. He earned his Bachelor's degree in chemistry from the University of Illinois in 1989, and he received his Ph.D. in chemistry from the University of Wisconsin, Madison, in 1993. Phil started at Abbott Laboratories as a postdoctoral fellow under Dr. Stephen Fesik and was converted to Staff Scientist in 1996. In his current role, Phil directs a multidisciplinary approach to lead generation and optimization and continues to apply and expand the utility of cheminformatic and fragment-based approaches to drug design.

Note Added after ASAP Publication. This manuscript was published on March 24, 2010 with an incorrect source cited in Table 3. The revised version was published on April 9, 2010.

References

(1) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44* (1), 235–249.
(2) Hughes, J. D.; Blagg, J.; Price, D. A.; Bailey, S.; Decrescenzo, G. A.; Devraj, R. V.; Ellsworth, E.; Fobian, Y. M.; Gibbs, M. E.; Gilles, R. W.; Greene, N.; Huang, E.; Krieger-Burke, T.; Loesel, J.; Wager, T.; Whiteley, L.; Zhang, Y. Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorg. Med. Chem. Lett.* **2008**, *18* (17), 4872–4875.
(3) Dong, X.; Gilbert, K. E.; Guha, R.; Heiland, R.; Kim, J.; Pierce, M. E.; Fox, G. C.; Wild, D. J. Web service infrastructure for chemoinformatics. *J. Chem. Inf. Model.* **2007**, *47* (4), 1303–1307.
(4) Ertl, P.; Muhlbacher, J.; Rohde, B.; Selzer, P. Web-based cheminformatics and molecular property prediction tools supporting drug design and development at Novartis. *SAR QSAR Environ. Res.* **2003**, *14* (5−6), 321–328.
(5) Hohman, M.; Gregory, K.; Chibale, K.; Smith, P. J.; Ekins, S.; Bunin, B. Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. *Drug Discovery Today* **2009**, *14* (5−6), 261–270.
(6) Schnecke, V.; Bostrom, J. Computational chemistry-driven decision making in lead generation. *Drug Discovery Today* **2006**, *11* (1−2), 43–50.
(7) Stahl, M.; Guba, W.; Kansy, M. Integrating molecular design resources within modern drug discovery research: the Roche experience. *Drug Discovery Today* **2006**, *11* (7−8), 326–333.
(8) Munos, B. Lessons from 60 years of pharmaceutical innovation. *Nat. Rev. Drug Discovery* **2009**, *8* (12), 959–968.
(9) Posner, B. A. High-throughput screening-driven lead discovery: meeting the challenges of finding new therapeutics. *Curr. Opin. Drug Discovery Dev.* **2005**, *8* (4), 487–494.
(10) Macarron, R. Critical review of the role of HTS in drug discovery. *Drug Discovery Today* **2006**, *11* (7−8), 277–279.
(11) Fox, S.; Farr-Jones, S.; Sopchak, L.; Boggs, A.; Nicely, H. W.; Khoury, R.; Biros, M. High-throughput screening: update on practices and success. *J. Biomol. Screen.* **2006**, *11* (7), 864–869.
(12) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A comparison of physiochemical property profiles of development and marketed oral drugs. *J. Med. Chem.* **2003**, *46* (7), 1250–1256.
(13) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45* (12), 2615–2623.
(14) Zhang, M. Q.; Wilkinson, B. Drug discovery beyond the "rule-of-five". *Curr. Opin. Biotechnol.* **2007**, *18* (6), 478–488.
(15) Lajiness, M. S.; Vieth, M.; Erickson, J. Molecular properties that influence oral drug-like behavior. *Curr. Opin. Drug Discovery Dev.* **2004**, *7* (4), 470–477.

(16) Vieth, M.; Sutherland, J. J. Dependence of molecular properties on proteomic family for marketed oral drugs. *J. Med. Chem.* **2006**, *49* (12), 3451–3453.
(17) Ganesan, A. The impact of natural products upon modern drug discovery. *Curr. Opin. Chem. Biol.* **2008**, *12* (3), 306–317.
(18) Lipinski, C. A. Chris Lipinski discusses life and chemistry after the Rule of Five. *Drug Discovery Today* **2003**, *8* (1), 12–16.
(19) Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat. Rev. Drug Discovery* **2007**, *6* (11), 881–890.
(20) Azzaoui, K.; Hamon, J.; Faller, B.; Whitebread, S.; Jacoby, E.; Bender, A.; Jenkins, J. L.; Urban, L. Modeling promiscuity based on in vitro safety pharmacology profiling data. *ChemMedChem* **2007**, *2* (6), 874–880.
(21) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96* (18), 9997–10002.
(22) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand efficiency: a useful metric for lead selection. *Drug Discovery Today* **2004**, *9* (10), 430–431.
(23) Leach, A. R.; Hann, M. M.; Burrows, J. N.; Griffen, E. J. Fragment screening: an introduction. *Mol. Biosyst.* **2006**, *2* (9), 430–446.
(24) Abad-Zapatero, C.; Metz, J. T. Ligand efficiency indices as guideposts for drug discovery. *Drug Discovery Today* **2005**, *10* (7), 464–469.
(25) Hajduk, P. J. Fragment-based drug design: How big is too big? *J. Med. Chem.* **2006**, *49*, 6972–6976.
(26) Congreve, M.; Chessari, G.; Tisi, D.; Woodhead, A. J. Recent developments in fragment-based drug discovery. *J. Med. Chem.* **2008**, *51* (13), 3661–3680.
(27) McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* **2002**, *45*, 1712–1722.
(28) Huth, J. R.; Mendoza, R.; Olejniczak, E. T.; Johnson, R. W.; Cothron, D. A.; Liu, Y.; Lerner, C. G.; Chen, J.; Hajduk, P. J. ALARM NMR: a rapid and robust experimental method to detect reactive false positives in biochemical screens. *J. Am. Chem. Soc.* **2005**, *127* (1), 217–224.
(29) Zhou, S.; Chan, E.; Duan, W.; Huang, M.; Chen, Y. Z. Drug bioactivation, covalent binding to target proteins and toxicity relevance. *Drug Metab. Rev.* **2005**, *37* (1), 41–213.
(30) Rishton, G. M. Reactive compounds and in vitro false positives in HTS. *Drug Discovery Today* **1997**, *2* (9), 382–384.
(31) Williams, D. P. Toxicophores: investigations in drug safety. *Toxicology* **2006**, *226* (1), 1–11.
(32) Williams, D. P.; Naisbitt, D. J. Toxicophores: groups and metabolic routes associated with increased safety risk. *Curr. Opin. Drug Discovery Dev.* **2002**, *5* (1), 104–115.
(33) Davis, A. M.; Keeling, D. J.; Steele, J.; Tomkinson, N. P.; Tinker, A. C. Components of successful lead generation. *Curr. Top. Med. Chem.* **2005**, *5* (4), 421–439.
(34) Bhal, S. K.; Kassam, K.; Peirson, I. G.; Pearl, G. M. The Rule of Five revisited: applying log D in place of log P in drug-likeness filters. *Mol. Pharm.* **2007**, *4* (4), 556–560.
(35) Johnson, T. W.; Dress, K. R.; Edwards, M. Using the Golden Triangle to optimize clearance and oral absorption. *Bioorg. Med. Chem. Lett.* **2009**, *19* (19), 5560–5564.
(36) Waring, M. J. Defining optimum lipophilicity and molecular weight ranges for drug candidates-Molecular weight dependent lower logD limits based on permeability. *Bioorg. Med. Chem. Lett.* **2009**, *19* (10), 2844–2851.
(37) Ryckmans, T.; Edwards, M. P.; Horne, V. A.; Correia, A. M.; Owen, D. R.; Thompson, L. R.; Tran, I.; Tutt, M. F.; Young, T. Rapid assessment of a novel series of selective CB(2) agonists using parallel synthesis protocols: A lipophilic efficiency (LipE) analysis. *Bioorg. Med. Chem. Lett.* **2009**, *19* (15), 4406–4409.
(38) Mannhold, R.; Poda, G. I.; Ostermann, C.; Tetko, I. V. Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96000 compounds. *J. Pharm. Sci.* **2009**, *98* (3), 861–893.
(39) Martin, Y. C. Let's not forget tautomers. *J. Comput.-Aided Mol. Des.* **2009**, *23* (10), 693–704.
(40) Jain, N.; Yalkowsky, S. H. Estimation of the aqueous solubility I: application to organic nonelectrolytes. *J. Pharm. Sci.* **2001**, *90* (2), 234–252.
(41) Jain, N.; Yang, G.; Machatha, S. G.; Yalkowsky, S. H. Estimation of the aqueous solubility of weak electrolytes. *Int. J. Pharm.* **2006**, *319* (1−2), 169–171.
(42) Delaney, J. S. Predicting aqueous solubility from structure. *Drug Discovery Today* **2005**, *10* (4), 289–295.
(43) Llinas, A.; Glen, R. C.; Goodman, J. M. Solubility challenge: can you predict solubilities of 32 molecules using a database of 100 reliable measurements? *J. Chem. Inf. Model.* **2008**, *48* (7), 1289–1303.
(44) Hewitt, M.; Cronin, M. T.; Enoch, S. J.; Madden, J. C.; Roberts, D. W.; Dearden, J. C. In silico prediction of aqueous solubility: the solubility challenge. *J. Chem. Inf. Model.* **2009**, *49* (11), 2572–2587.
(45) Sheridan, R. P. The most common chemical replacements in drug-like compounds. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (1), 103–108.
(46) Haubertin, D. Y.; Bruneau, P. A database of historically-observed chemical replacements. *J. Chem. Inf. Model.* **2007**, *47* (4), 1294–1302.
(47) Leach, A. G.; Jones, H. D.; Cosgrove, D. A.; Kenny, P. W.; Ruston, L.; MacFaul, P.; Wood, J. M.; Colclough, N.; Law, B. Matched molecular pairs as a guide in the optimization of pharmaceutical properties: a study of aqueous solubility, plasma protein binding and oral exposure. *J. Med. Chem.* **2006**, *49* (23), 6672–6682.
(48) Hajduk, P. J.; Sauer, D. R. Statistical analysis of the effects of common chemical substituents on ligand potency. *J. Med. Chem.* **2008**, *51* (3), 553–564.
(49) Gleeson, P.; Bravi, G.; Modi, S.; Lowe, D. ADMET rules of thumb II: A comparison of the effects of common substituents on a range of ADMET parameters. *Bioorg. Med. Chem.* **2009**, *17* (16), 5906–5919.
(50) Lewis, M. L.; Cucurull-Sanchez, L. Structural pairwise comparisons of HLM stability of phenyl derivatives: Introduction of the Pfizer metabolism index (PMI) and metabolism-lipophilicity efficiency (MLE). *J. Comput.-Aided Mol. Des.* **2009**, *23* (2), 97–103.
(51) Lovering, F.; Bikker, J.; Humblet, C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **2009**, *52* (21), 6752–6756.
(52) Ritchie, T. J.; Macdonald, S. J. The impact of aromatic ring count on compound developability—are too many aromatic rings a liability in drug design? *Drug Discovery Today* **2009**, *14* (21−22), 1011–1020.
(53) Hajduk, P. J.; Mendoza, R.; Petros, A. M.; Huth, J. R.; Bures, M.; Fesik, S. W.; Martin, Y. C. Ligand binding to domain-3 of human serum albumin: a chemometric analysis. *J. Comput.-Aided Mol. Des.* **2003**, *17* (2−4), 93–102.
(54) Stewart, K. D.; Shiroda, M.; James, C. A. Drug Guru: a computer software program for drug design using medicinal chemistry rules. *Bioorg. Med. Chem.* **2006**, *14* (20), 7011–7022.
(55) Mao, H.; Hajduk, P. J.; Craig, R.; Bell, R.; Borre, T.; Fesik, S. W. Rational design of diflunisal analogues with reduced affinity for human serum albumin. *J. Am. Chem. Soc.* **2001**, *123* (43), 10429–10435.
(56) Wendt, M. D.; Shen, W.; Kunzer, A.; McClellan, W. J.; Bruncko, M.; Oost, T. K.; Ding, H.; Joseph, M. K.; Zhang, H.; Nimmer, P. M.; Ng, S. C.; Shoemaker, A. R.; Petros, A. M.; Oleksijew, A.; Marsh, K.; Bauch, J.; Oltersdorf, T.; Belli, B. A.; Martineau, D.; Fesik, S. W.; Rosenberg, S. H.; Elmore, S. W. Discovery and structure-activity relationship of antagonists of B-cell lymphoma 2 family proteins with chemopotentiation activity in vitro and in vivo. *J. Med. Chem.* **2006**, *49* (3), 1165–1181.
(57) Huynh, L.; Masereeuw, R.; Friedberg, T.; Ingelman-Sundberg, M.; Manivet, P. In silico platform for xenobiotics ADME-T pharmacological properties modeling and prediction. Part I: Beyond the reduction of animal model use. *Drug Discovery Today* **2009**, *14* (7−8), 401–405.
(58) Jacob, A.; Pratuangdejkul, J.; Buffet, S.; Launay, J. M.; Manivet, P. In silico platform for xenobiotics ADME-T pharmacological properties modeling and prediction. Part II: The body in a Hilbertian space. *Drug Discovery Today* **2009**, *14* (7−8), 406–412.
(59) Wang, J.; Hou, T. Recent advances on in silico ADME modeling. *Ann. Rep. Comp. Chem.* **2009**, *5*, 101–123.
(60) Valerio, L. G. In silico toxicology for the pharmaceutical sciences. *Toxicol. Appl. Pharmacol.* **2009**, *241* (3), 356–370.
(61) Kelder, J.; Grootenhuis, P. D.; Bayada, D. M.; Delbressine, L. P.; Ploemen, J. P. Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.* **1999**, *16* (10), 1514–1519.
(62) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43* (20), 3714–3717.
(63) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45* (19), 4350–4358.
(64) Leo, A.; Weininger, A. *Daylight Chemical Information Systems, 3*; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA, 1995.

(65) Rogers, D.; Brown, R. D.; Hahn, M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen.* **2005**, *10* (7), 682–686.
(66) Whittle, M.; Gillet, V. J.; Willett, P.; Loesel, J. Analysis of data fusion methods in virtual screening: similarity and group fusion. *J. Chem. Inf. Model.* **2006**, *46* (6), 2206–2219.
(67) Baber, J. C.; Shirley, W. A.; Gao, Y.; Feher, M. The use of consensus scoring in ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46* (1), 277–288.
(68) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *J. Chem. Inf. Model.* **2008**, *48* (5), 941–948.
(69) Rush, T. S., 3rd; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein−protein interaction. *J. Med. Chem.* **2005**, *48* (5), 1489–1495.
(70) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46* (4), 499–511.
(71) Ertl, P. In silico identification of bioisosteric functional groups. *Curr. Opin. Drug Discovery Dev.* **2007**, *10* (3), 281–288.
(72) Boda, K.; Seidel, T.; Gasteiger, J. Structure and reaction based evaluation of synthetic accessibility. *J. Comput.-Aided Mol. Des.* **2007**, *21* (6), 311–325.
(73) Todd, M. H. Computer-aided organic synthesis. *Chem. Soc. Rev.* **2005**, *34* (3), 247–266.
(74) Hare, B. J.; Walters, W. P.; Caron, P. R.; Bemis, G. W. CORES: an automated method for generating three-dimensional models of protein/ligand complexes. *J. Med. Chem.* **2004**, *47* (19), 4731–4740.
(75) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular shape and medicinal chemistry: a perspective. *J. Med. Chem.* **2010**, in press.
(76) Doweyko, A. M. Is QSAR relevant to drug discovery? *IDrugs* **2008**, *11* (12), 894–899.
(77) Doweyko, A. M. QSAR: dead or alive? *J. Comput.-Aided Mol. Des.* **2008**, *22* (2), 81–89.
(78) Paolini, G. V.; Shapland, R. H.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24* (7), 805–815.
(79) Keiser, M. J.; Setola, V.; Irwin, J. J.; Laggner, C.; Abbas, A. I.; Hufeisen, S. J.; Jensen, N. H.; Kuijer, M. B.; Matos, R. C.; Tran, T. B.; Whaley, R.; Glennon, R. A.; Hert, J.; Thomas, K. L.; Edwards, D. D.; Shoichet, B. K.; Roth, B. L. Predicting new molecular targets for known drugs. *Nature* **2009**, *462* (7270), 175–181.
(80) Ekins, S.; Nikolsky, Y.; Nikolskaya, T. Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity. *Trends Pharmacol. Sci.* **2005**, *26* (4), 202–209.
(81) Krishna, R.; Schaefer, H. G.; Bjerrum, O. J. Effective integration of systems biology, biomarkers, biosimulation, and modeling in streamlining drug development. *J. Clin. Pharmacol.* **2007**, *47* (6), 738–743.
(82) Michelson, S. The impact of systems biology and biosimulation on drug discovery and development. *Mol. Biosyst.* **2006**, *2* (6−7), 288–291.
(83) Michelson, S.; Cole, M. The future of predictive biosimulation in drug discovery. *Expert Opin. Drug Discovery* **2007**, *2* (4), 515–523.
(84) Grant, J. A.; Pickup, B. T.; Sykes, M. J.; Kitchen, C. A.; Nicholls, A. The Gaussian generalized Born model: application to small molecules. *Phys. Chem. Chem. Phys.* **2007**, *9* (35), 4913–4922.
(85) Leo, A. *CLOGP*, 4.0; Biobyte Corporation: Claremont CA, 2002.
(86) Lipinski, C.; Lombardo, F.; Dominy, B.; Feeney, P. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.
(87) ACD PhysChem Batch Software, www.acdlabs.com/physchem-batch.
(88) Bruneau, P.; McElroy, N. R. logD7.4 modeling using Bayesian regularized neural networks. Assessment and correction of the errors of prediction. *J. Chem. Inf. Model.* **2006**, *46* (3), 1379–1387.