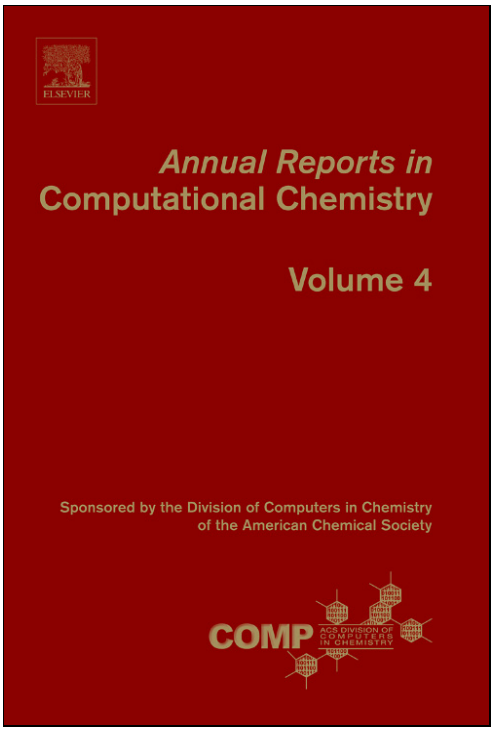


From Alan C. Cheng, Predicting Selectivity and Druggability in Drug Discovery. In: Ralph A. Wheeler and David C. Spellmeyer, editors, Annual Reports in Computational Chemistry, Volume 4. Amsterdam: Elsevier, 2008, p. 23
ISBN: 978-0-444-53250-3

Elsevier

CHAPTER 2

Predicting Selectivity and Druggability in Drug Discovery

Alan C. Cheng*

Contents

1. Introduction 23
2. Selectivity 24
   2.1. Ligand analysis 24
   2.2. Sequence analysis 25
   2.3. Structure-based analysis 26
3. Druggability 29
   3.1. Ligand analysis 29
   3.2. Sequence analysis 30
   3.3. Structure analysis 30
4. Conclusions 33
References 33

1. INTRODUCTION

Druggability and selectivity analysis are increasingly performed in early drug discovery for both target assessment and setting lead optimization strategies. This is necessitated by the high failure rates in the drug discovery process—greater than 60% in early drug discovery screening and lead optimization stages alone [1]. In target assessment, ideas are rated on target-validation, assay feasibility, druggability, and selectivity as it relates to toxicity and side-effect potential [2,3]. In lead optimization, selectivity analysis can suggest both possible selectivity issues, as well as regions of the binding site that allow the drug discovery team to overcome these issues. Druggability analysis can be useful in suggesting additional "hot spots" for increasing potency of lead compounds. This review covers computational approaches for assessing and predicting selectivity and druggability, and those interested in computational aspects of target validation may want to read a recent review by Loging et al. [4].

* Amgen Inc., One Kendall Square, Bldg 1000, Cambridge, MA 02139, USA. E-mail: alan.cheng@amgen.com

Annual Reports in Computational Chemistry, Vol. 4
ISSN 1574-1400, DOI: 10.1016/S1574-1400(08)00002-9

2. SELECTIVITY

Selectivity analysis has the goal of identifying potential secondary pharmacology and suggesting assays for following up predicted selectivity issues, as well as identifying strategies for improving selectivity of small molecule leads. Profiling for and optimizing against secondary pharmacology is important to the discovery of compounds with decreased side effects, more desirable therapeutic profiles, and greater therapeutic differentiation. For example, successful kinase inhibitors such as imantinib (Gleevec) and sutanimib (Sutent) have a distinct selectivity profile that confers efficacy and safety [5]. In terms of computational approaches, analyses of increasing sophistication can be performed depending on how much information is available—this includes protein sequence, protein structure, and ligand information. The wide availability of these types of information for protein kinases has allowed for a significant body of selectivity analysis work, which is covered in reviews such as [6–9]. Readers interested in off-target cytochrome P450 inhibition, transporter-mediated efflux, and ADMET-related prediction for small molecules may want to consider recent reviews in the Annual Reports in Computational Chemistry [10–12].

2.1 Ligand analysis

If ligands are known for the biological target, cheminformatics approaches are useful in identifying potential selectivity issues, especially when they are used in combination with aggregate compound databases that are annotated with biological activity. Such databases include historical databases maintained in-house at biopharmaceutical companies, as well as Jubilant, GVK, MDDR, WOMBAT, and StARLITe databases [13,14]. The idea of comparing targets by looking at the small molecules that modulate them has been termed SARAH, for structure-activity relationship homology [15]. In the original SARAH idea, experimentally measured affinities for a diverse set of compounds represent an "affinity fingerprint" for a target, and similar pharmacological profiles would indicate target homology in SAR space [15]. From a drug discovery perspective, this approach is meaningful since it identifies similarity based on inhibitor or antagonist/agonist profiles, and an example where this experimental approach is applied to cysteine protease inhibitors is described in [16].

Small molecule screening data accumulated in compound databases can also be used in identifying target homology in SAR space. One approach is to identify analogs of the known active compounds using similarity searches based on 2D chemical fingerprints [17], and then look at the biological activities of the identified compounds. Conceptually, the confidence in the predicted selectivity issue increases as a pair of biological targets share larger numbers of chemotypes. This approach has traditionally been qualitative and subjective, and recently several groups have sought to make the approach more systematic and rigorous [18]. One way, termed the similarity ensemble approach (SEA) [19], takes the summed similarity score over all pairs of ligands that two biological targets share and compares it to the distribution of scores from random sets of compounds, thus allowing calculation of a statistical confidence value that is similar

to the E-value used in scoring BLAST [20] sequence searches. Another approach is to generate a similarity score between two sets of ligands, but use Bayesian models to weight compound substructures that contribute more to activity [21]. A pure machine learning approach can also be used, and involves training activity models and then using the models to predict off-target activities. Nidhi et al. used a naïve Bayesian classifier to train models for 964 biological target activities and found 77% prediction accuracy when predicting activities for a separate data set [22].

These chemo-centric similarity approaches can help in identifying a selectivity panel if there is sufficient *a priori* data, and can also be used after a high-throughput screen (HTS) is complete and more ligands are known. The database SARAH approach has been applied in varying degrees of sophistication to nuclear hormone receptors [23], kinases [8], and enzymes in general [24].

2.2 Sequence analysis

While ligand information is not always available, protein sequence information is almost always available. Starting with the protein sequence, related proteins, or homologs, can be found through sequence similarity searches such as BLAST [20], where the typical search is a protein BLAST against human sequences in the non-redundant sequence database [25]. Rat and mouse sequences may also be of interest depending on the disease model that will be used. Once protein sequences are identified, multiple sequence alignment of significant BLAST hits can be performed using programs such as Clustal [26].

A variety of methods are available to cluster sequences and identify similarities starting from the multiple sequence alignment, with the most straightforward of these being pair-wise measurement of sequence identity or sequence similarity. Sequence identity is the percent of the residue positions that match, while sequence similarity involves a substitution matrix where amino acid residue similarity is taken into account. A more sophisticated approach to identify significantly related proteins is to infer a phylogenetic tree based on the multiple sequence alignment. Closely related proteins in a phylogenetic tree are, in general, likely to be selectivity issues. Similarity and phylogenetic tree calculations can be performed using software tools such as PFAAT [27], Jalview [28], and Mega [29], which are listed in Table 2.1.

When information on the protein domain of interest is available, the sequence analysis can be focused on the domain sequences. An even more detailed investigation of residues around the binding site can me made if there is information about the desired drug interaction site from experimental data such as mutagenesis results or co-crystal structure information. For instance, in analyzing kinase selectivity issues, workers often focus on residues lining the ATP binding site [30,31]. In these binding site analyses, it is important to note that all residues do not contribute equally to binding, and close inspection of the actual interactions based on a crystal structure would be prudent [32,33]. For instance, a protein backbone interaction to the ligand does not depend strongly on amino acid type, and prolines can change or rigidify the main chain conformation. An analysis of ki-

Table 2.1 Some popular tools for performing selectivity analysis using protein sequence

Task	Tool/resource	Web link
Search for related protein sequences	Blast	http://www.ncbi.nlm.nih.gov/blast
Multiple sequence alignment	Clustal	http://bips.u-strasbg.fr/en/Documentation/ClustalX/
Analyze a multiple sequence alignment	PFAAT Jalview Mega	http://pfaat.sourceforge.net http://www.jalview.org http://www.megasoftware.net
Identifying domains	NCBI conserved domain database	http://www.ncbi.nlm.nih.gov/Structure/cdd/

nase inhibitors found that two non-conservative, energetically-important, residue substitutions in the binding site are sufficient for gaining selectivity for a compound [34].

When co-crystal structures are not available, Ortiz et al. have suggested using functional residue prediction methods to identify selectivity residues [35]. The most popular of these methods are Evolutionary Trace [36] and ConSurf [37], which use phylogenetic trees to predict biologically-relevant residues that are then mapped onto a representative crystal structure.

2.3 Structure-based analysis

One significant limitation of sequence-based approaches is the inability to assess selectivity issues between targets lacking sequence homology. For instance, protein kinase sequences cannot be aligned to phosphodiesterase sequences even though selectivity issues have been observed between the two target classes. Structure-based approaches can help to identify non-homologous selectivity concerns when co-crystal structures are available. Such approaches can be classified by whether they are receptor-focused or ligand-focused, as described below.

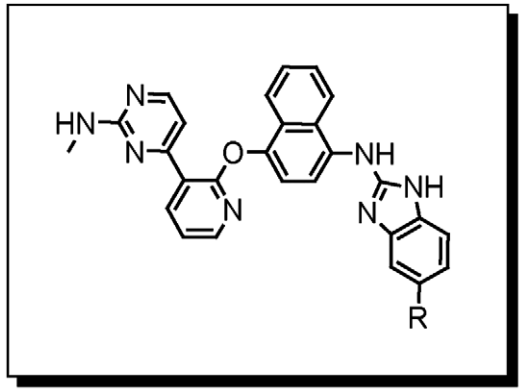
Receptor-focused approaches involve comparison of the physiochemical properties of residues that line the binding pocket. CavBase [38], SURFACE [39], and SitesBase [40] are three examples. CavBase converts the portions of binding site residues exposed to solvent into sets of points defined by one of five pseudocenter types (aliphatic, donor, acceptor, donor/acceptor, and aromatic). For instance, a tyrosine is represented by an "aromatic" pseudocenter placed in the middle of the phenyl ring and a "donor/acceptor" pseudocenter placed at the oxygen of the hydroxyl. The constellation of pseudocenters representing a binding site is then compared to those of other binding sites using clique-detection algorithms that identify matching portions of the constellations. This approach has been applied

to classification of the enzyme binding pocket in protein kinases [41]. SURFACE represents each residue using just two pseudocenters, which represents the backbone Cα atom and the side-chain center of mass [39]. Instead of just scoring for matches, an evolutionary amino acid substitution matrix is used. SitesBase compares binding sites based on actual atoms and atom types (carbon, nitrogen, oxygen, sulfur) instead of pseudocenters [40], and the approach was applied to proteases [42].

Instead of basing comparisons off properties of residues that flank the binding site, ligand-focused approaches attempt to compare the actual small molecule binding space. GRID/PCA [43,44], for instance, uses GRID to systematically sample the binding site with a set of chemical probes, and uses an energy function to generate molecular interaction fields that represent areas of favorable affinity for each of the probes. Applying principle component analysis (PCA) to the GRID values then identifies consistency as well as differences in the interaction fields. The method has been used to study a set of 13 ephrin receptor tyrosine kinases [45] as well as a set of ten structures of CDK2 and GSK-3b [46]. Reported applications of GRID/PCA have generally focused on selectivity issues among proteins with sequence homology, in part due to the necessity of receptor structure superposition. The more recent GRIND/PCA method [43] does allow for comparisons independent of structural alignment, and has been used to compare a homology model of adenosine receptor A1 to four ribose-binding proteins [38]. The small number of comparisons is likely due to the compute-intensive nature of GRID calculations. Another approach [47] uses docking to identify a set of predicted active compounds for the protein of interest, and this set is then docked to possible selectivity targets. Targets with the most similar binding sites were shown to have the highest docking scores.

For lead optimization, Sheinerman et al. [34] showed in the context of protein kinases that energetically important residues could be identified using a systematic analysis of small-molecule structure-activity relationships in the context of a protein family sequence alignment and available structures for compound binding modes. A quantitative method for optimization of electrostatic interactions—including accounting for desolvation effects—was demonstrated for HIV protease inhibitor design recently [48]. The approach uses mathematical optimization techniques to define a ligand with maximal potency against a desirable set of targets (a set of escape mutants for HIV protease) and minimal potency for an undesirable set of human aspartyl protease "decoys" [48]. A rigorous but decidedly theoretical biophysical inquiry into the physical basis of selectivity found that polar and charged groups increase specificity of ligand interactions due to their greater sensitivity to shape complementarity as compared to hydrophobic interactions, and, in addition, conformational flexibility can increase the specificity of polar and charged interactions [49].

Molecular analyses are imperfect in that serendipitous binding modes are always possible. For instance, crystal structures of PXR show that ligands have multiple binding modes [50], and protein kinases can adopt multiple inactive conformations that are druggable [51]. A dramatic example is shown in Figure 2.1, where a small modification to a non-selective kinase inhibitor yielded 1400× se-



R=H	R=CF_3
Tie-2: IC_{50}=30 nM	Tie-2: IC_{50}=1 nM
KDR: IC =70 nM	KDR: IC =1,400 nM

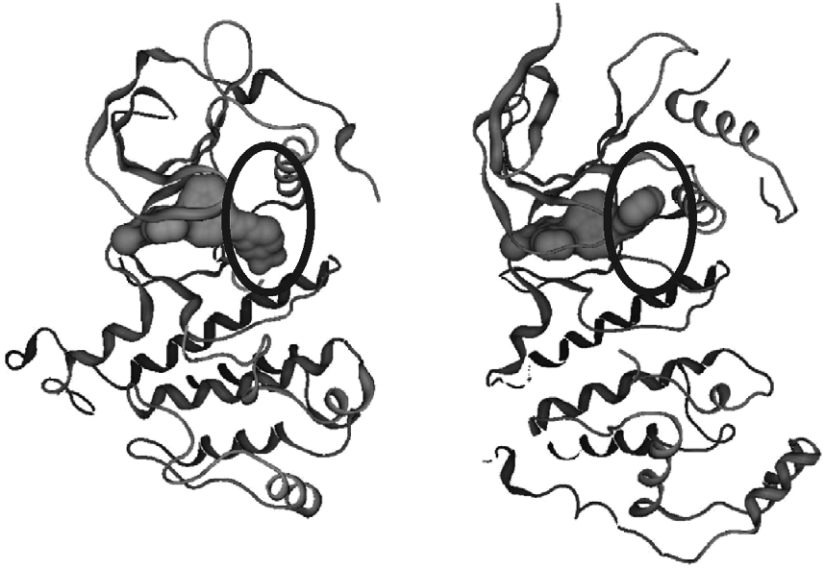


FIGURE 2.1

lectivity, most likely because of a flip in the binding of a terminal benzimidazole group [52]. However, in protein kinases, most, if not all, of the major binding modes have probably been identified and can be used in selectivity analyses [51], although there is a possibility that allosteric sites away from the ATP active site exist [53]. While serendipitous binding modes are an infrequent but important consideration, computational methods are nevertheless useful as systematic, objective analyses for assessing the risk of selectivity issues as well as identifying possible selectivity issues and strategies that should be experimentally considered.

3. DRUGGABILITY

What is "druggability"? It is ultimately the success of the compound in human clinical trials. This includes not only compound properties but also aspects of efficacy, safety, and commercial attractiveness which are difficult to predict. For scientists engaged in drug discovery prior to clinical trials, 'druggability' can be defined more tangibly in terms of the chemical matter at the high-throughput screening and lead generation stages.

3.1 Ligand analysis

Traditionally, druggability has been assessed experimentally. At the HTS stage, teams typically define "druggability" in terms of identifying a "druglike" small molecule with activity in the one micromolar range, where the term "druglike" refers to compounds with physical properties ranges similar to known oral drugs [54–57]. Common "druglike" rules include polar surface area (PSA) less than 140 A^2 [55], number of rotateable bonds less than 10 [56], molecular weight less than about 500 Da, and no more than one rule violation in the Lipinski Rule-of-Five [54]. For a more complete review of "druglike" properties, please see the recent Annual Reports in Computational Chemistry review [58]. Project teams may preferentially identify "leadlike" compounds with lower molecular weights and ClogP values [57]. In the next stage, the lead generation stage, the project team typically defines druggability as the potential to find a compound with nanomolar potency, drug-like properties, as well as experimentally-measured properties related to unwanted secondary pharmacology (for example, selectivity in the CEREP or MDS Panlabs panel), metabolism (microsomal stability, hepatocyte stability, and cytochrome P450 inhibition), and intestinal absorption (permeability, rodent pharmacokinetics).

A group at Abbott has demonstrated that hit-rates from NMR-based fragment screening are a good indicator of the target's druggability [59,60]. The fragment library consisted of "fragmentlike" compounds that have an average molecular weight of 220 and average ClogP of 1.5. Screening a fragment library of around 10,000 compounds using NMR technologies may be more cost effective than screening a full compound library that commonly contain over a million compounds.

3.2 Sequence analysis

In addition to the largely experimental screening approaches, druggability can be assessed based on bioinformatics analysis of the protein sequence. Sequence similarity can be used to determine whether the gene of interest is part of a gene family or sub-family with known druggability status [61]. For instance, aminergic G-protein-coupled receptors (GPCR) and protein kinases are known to be druggable based on marketed drugs as well as collective HTS and medicinal chemistry experience, and so a new aminergic GPCR or protein kinase would be expected to be druggable as well. Hopkins and Groom did a systematic "druggable genome" analysis to identify 130 gene families that are targeted by rule-of-five compliant compounds, and they then identified proteins from the human genome that map to these gene families [61]. The results suggest that only 10% of genes in the human genome map to precedented druggable gene families, and that only 5% are both druggable and disease-relevant. The analysis has been updated by Overington et al. [62] as well as others [63–66], and workers at Novartis have set up a public web server for running a target sequence query against known druggable sequences at http://function.gnf.org/druggable/index.html, although the server is only for academic and non-profit use [66]. Some have pointed out that the much larger "druggable proteome" or "druggable targetome" is more relevant than the "druggable genome" [67]. For instance, the proteasome can be inhibited by a small molecule, and, in addition, there is emerging evidence that protein–protein complexes such as MDM2-p53 are druggable. Nevertheless, the argument that only a small fraction of targets are druggable is not generally contested, and argues for the importance of assessing targets systematically [2,3].

3.3 Structure analysis

Whereas small molecule drugs usually bind to pockets, the reverse is not always true—not all pockets on a biological target are druggable. Upon inspecting crystal structures of druggable and difficult druggability protein binding sites, it becomes clear that druggable pockets tend to be deep, hydrophobic, and of a limited size. Druggable pockets tend to reflect the properties of the drug-like ligands that they bind, and so they might also be called "drug-like binding sites" or "beautiful binding sites" [61,68].

How does one identify beautiful binding sites? Available algorithms include those for identifying ligand-binding "hot spots" on the surface of protein structures, which include fragment-based approaches as well as statistical approaches based on structural descriptors. A computational solvent mapping approach was able to identify known druggable pockets based on known crystal structures, and can further be used to identify hotspots on protein surfaces [69]. Another approach precalculates a van der Waals potential at nodes of a grid that envelops the protein, and then searches for high-scoring grid clusters in order to predict ligand binding pockets [70,71]. Statistical learning approaches include those developed for identifying functional sites [72]. One example is a neural net approach called HotPatch [73] that is based on calculated electrostatic potential, charge, concavity, surface roughness, and hydrophobicity values. HotPatch was successfully

used by another group to predict an allosteric small-molecule site in caspases [74]. Another algorithmic approach that is easier to interpret combines probabilistic distribution functions (PDFs) for a similar set of properties [75]. SiteMap (Schrodinger, Inc.) identifies and scores binding sites based on the typical physiochemical properties and, additionally, a "hydrophobic enclosure" term, which accounts for pocket shape in hydrophobic desolvation [76,77].

The approaches discussed so far, however, are focused on predicting pockets for any ligand, as opposed to predicting sites for *druglike* ligands or how druggable a given binding site is. The statistical approaches discussed so far could, in theory, capture druggability given a training set. Work by Hajduk et al. [59] was the first published approach using a statistical method to directly address druggability prediction. They derived a druggability scoring function by performing statistical regression of physiochemical properties calculated for a variety of protein pockets to hit-rates from NMR screening of "leadlike" fragments. More specifically, protein pockets defined using an InsightII (Accelrys, Inc.) flood-fill algorithm were analyzed to generate physiochemical descriptors, including surface area, volume, roughness, and number of charged residues, as well as descriptors of the pocket shape—pocket compactness and three principal moment descriptors. These calculated descriptors were then fitted to NMR fragment-screening data to yield a score, termed the 'druggability index' (D_I):

$$\begin{aligned} \textit{Druggability index} = {} & -14.0 \cdot X_{PocketCompactness} + 13.6 \cdot \log(X_{PocketCompactness}) \\ & + 2.98 \cdot \log(X_{ApolarContactArea}) - 0.023 \cdot X_{ApolarContactArea} \\ & + 2.98 \cdot \log(X_{SurfaceArea}) - 0.44 \cdot \log(X_{PolarContactArea}) \\ & + 1.2 \cdot \log(X_{ThirdPrincipalMoment}) - 1.03 \cdot \log(X_{FirstPrincipalMoment}) \\ & + 0.71 \cdot X_{Roughness} \\ & - 0.16 \cdot X_{NumberChargedResidues} \\ & - 1.11. \end{aligned}$$

Other descriptors included in the regression (volume, polar surface area, total contact area, and second principal moment) were found to be insignificant and not included in the final equation. With the training set of 23 proteins, the model yielded an r^2 of 0.65 and a q^2 of 0.56. On an external test set of 35 proteins, 94% of the known druggable pockets were correctly predicted as druggable.

In another work, Cheng et al. took an approach that combines a biophysics model with the concept of drug-like physiochemical properties [78]. Intuitively, druggable pockets are hydrophobic [68,79], deep, and have a limited size. The authors used a literature biophysical model for the hydrophobic effect and normalized the equation for drug-like size. The resulting 'Maximal Affinity Prediction' (MAP) equation is an estimate of the maximal affinity of a given binding site for a small molecule with 'druglike' properties.

The MAP score is a continuous score reported as the estimated best K_d achievable by a passively absorbed, non-covalent oral drug:

$$\textit{Maximal drug-like affinity} = -\gamma(r) \cdot \frac{A_{nonpolar}^{target}}{A_{total}^{target}} \cdot 300\ \text{Å}^2, \quad \text{where } \gamma(r) = \frac{45 \frac{cal}{mol \cdot \text{Å}^2}}{1 - \frac{1.4}{r_{curvature}}}.$$

The surface areas, A, are the measured nonpolar and total surface areas on the defined binding pocket, and, for a concave pocket, the $\gamma(r)$ term represents how easily water will leave a hydrophobic cavity [80–83]. A deeper pocket would have a smaller radius of curvature, $r_{curvature}$, and thus a larger $\gamma(r)$, indicating that water will leave more easily, while a completely flat surface would have an $r_{curvature}$ of infinity. The model is based on a physical model describing hydrophobic free energies of hydrocarbons in water [80,82], from which the authors then normalized the surface area to account for drug-like properties—in particular, they normalized for a druglike molecular weight cut-off of about 550 Da, which is equivalent to about 300 Å^2 of surface area [78]. Interestingly, drug-like PSA constraints (<140 Å^2) are accounted for in the model if we assume that the protein pocket PSA complements the ligand PSA. A high polar surface area on the protein pocket will reduce the predicted maximal druglike affinity, and for a pocket with PSA = 140 Å^2 and fairly deep $r_{curvature} = 6$ Å, the MAP score is 5 μM.

In the MAP model, druggable targets have predicted K_d's in the nM range, while difficult targets had predicted K_d's greater than 100 nM. In the retrospective analysis, several targets were predicted to be difficult targets despite drugs being on the market. Through scholarship they found that these predicted difficult targets were only druggable through a prodrug or active transport approach, pointing out that the approach is useful for predicting passively-absorbed, oral druggability, and other approaches for achieving druggability such as covalent adduct formation, metal chelation, prodrug development, active transport, and allosteric modulation should be kept in mind for difficult targets. In a forward prediction experiment, the authors successfully predicted the druggability of two novel drug targets, fungal homoserine dehydrogenase and haemopoetic prostaglandin D synthase, where druggability was determined by the outcome of a high-throughput screen and subsequent lead optimization at Pfizer [78]. In theory, the lower the maximal affinity of the target binding site, the more freedom the team has in modifying the compound to optimize pharmacodynamic and pharmacokinetic properties while maintaining efficacious potencies. The druggability boundary of 100 nM is generic, and the quantitative predicted K_d values can be useful where a nM affinity inhibitor is, based on physiology, not needed [84].

Although the D_I and MAP equations have different forms, the dominant terms are exceptionally consistent. In the MAP model non-polar surface area and curvature are the properties used, while in the D_I model the highest-weighted descriptors are surface area and pocket compactness (which correlates with curvature).

Since the methods use static crystal structures, one natural issue is how to capture protein flexibility. The MAP approach was robust to differences in the binding site between different co-crystal structures for a set of enzymes where multiple co-crystals were available at the time. The binding sites, however, were all enzyme

binding sites that largely consist of stable secondary structure motifs (helices and sheets), and binding sites that are composed of long unstructured regions will certainly see larger variations. For more flexible binding sites, molecular dynamics (MD) simulation could be used. Indeed, applying the D_I approach to snapshots from a MD simulation was necessary and sufficient to correctly predict the druggability of FKBP, Bcl-xL, and AKT-PH domain small-molecule binding sites [85]. These three binding sites involve long loop regions. Another group showed that for three protein-protein binding sites (Bcl-xl, IL-2, and MDM2), MD simulations starting from the apo-crystal conformations successfully resulted in sampling of the known small-molecule bound conformation [86]. Both studies use known druggable proteins, and it would be useful to know if difficult targets can be correctly assessed as well. The lesson here might be that even when starting from ligand-bound structures, care should be taken in assessing druggability of binding sites involving any loop regions, and static structures should not be used at all without simulation of their flexibility if the binding site is formed partially by loop regions. If the experimental conformation in hand is calculated to be sufficiently druggable however, flexibility becomes a non-issue. Flexible protein surfaces can reveal more druggable binding sites than the static structures indicate, as been shown crystallographically in the case of IL-2 [87,88], and calculating the inherent flexibility or adaptability of a site may help in predicting its druggability [89,90].

For targets assessed or found through experience to be difficult, computational methods can aid in lead optimization. In general, structure-based design methods, such as those discussed earlier for identifying hot spots as well as those reviewed in [91], can be useful in driving potency in a more directed manner. Pockets that are difficult to drug tend to be polar, and quantitative charge-optimization approaches can be useful in optimizing leads based on electrostatic interactions, taking into account ligand and receptor desolvation which can be difficult to visualize [92]. Allosteric modulation is increasingly sought [93], and emerging computational methods that combine druggable pocket prediction with functional residue prediction may eventually aid allosteric drug identification [53].

4. CONCLUSIONS

This review covered cheminformatics, bioinformatics, and structure-based drug design approaches and how they aid assessment of selectivity and druggability as well as setting of lead optimization strategies. While computational approaches will continue to improve in accuracy, they are nevertheless useful today for bringing together data in a rational, model-based manner to inform experiments and decision making.

REFERENCES

1. Brown, D., Superti-Furga, G. Rediscovering the sweet spot in drug discovery. Drug Discov. Today 2003, 8, 1067–77.

2. Stahl, M., Guba, W., Kansy, M. Integrating molecular design resources within modern drug discovery research: The Roche experience. Drug Discov. Today 2006, 11, 326–33.
3. Frearson, J.A., Wyatta, P.G., Gilberta, I.H., Fairlamba, A.H. Target assessment for antiparasitic drug discovery. Trends Parasitol. 2007, 23, 589–95.
4. Loging, W., Harland, L., Williams-Jones, B. High-throughput electronic biology: Mining information for drug discovery. Nature Rev. Drug. Discov. 2007, 6, 220–30.
5. Daub, H., Specht, K., Ullrich, A. Strategies to overcome resistance to targeted protein kinase inhibitors. Nature Rev. Drug Disc. 2004, 3, 1001–10.
6. Knight, Z.A., Shokat, K.M. Features of selective kinase inhibitors. Chem. Biol. 2005, 12, 621–37.
7. Rockey, W.M., Elcock, A.H. Rapid computational identification of the targets of protein kinase inhibitors. J. Med. Chem. 2005, 48, 4138–52.
8. Vieth, M., Higgs, R.E., Robertson, D.H., Shapiro, M., Gragg, E.A., Hemmerle, H. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. Biochim. Biophys. Acta 2004, 1697, 243–57.
9. Vieth, M., Sutherland, J.J., Robertson, D.H., Campbell, R.M. Kinomics: Characterizing the therapeutically validated kinase space. Drug Discov. Today 2005, 10, 839–46.
10. Fox, T., Kriegl, J.M. Linear quantitative structure–activity relationships for the interaction of small molecules with human cytochrome P450 isoenzymes. Ann. Reports Comp. Chem. 2005, 1, 63–81.
11. Verras, A., Kuntz, I.D., Ortiz de Montellano, P.R. Cytochrome P450 enzymes: Computational approaches to substrate prediction. Ann. Reports Comp. Chem. 2005, 1, 171–95.
12. Clark, D.E. Computational prediction of ADMET properties: Recent developments and future challenges. Ann. Reports Comp. Chem. 2005, 1, 133–51.
13. Oprea, T.I., Tropsha, A. Target, chemical and bioactivity databases—Integration is key. Drug Disc. Today: Technologies 2006, 3, 357–66.
14. Nidhi, Glick, M., Davies, J.W., Jenkins, J.L. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. J. Chem. Inf. Model. 2006, 46, 1124–33.
15. Frye, S.V. Structure-activity relationship homology (SARAH): A conceptual framework for drug discovery in the genomic era. Chem Biol. 1999, 6, R3–7.
16. Greenbaum, D.C., Arnold, W.D., Lu, F., Hayrapetian, L., Baruch, A., Krumrine, J., Toba, S., Chehade, K., Brömme, D., Kuntz, I.D., Bogyo, M. Small molecule affinity fingerprinting. A tool for enzyme family subclassification, target identification, and inhibitor design. Chem. Biol. 2002, 9, 1085–94.
17. Martin, Y.C., Kofron, J.L., Traphagen, L.M. Do structurally similar molecules have similar biological activity? J. Med. Chem. 2002, 45, 4350–8.
18. Paolini, G.V., Shapland, R.H.B., van Hoorn, W.P., Mason, J.S., Hopkins, A.L. Global mapping of pharmacological space. Nature Biotech. 2006, 24, 805–15.
19. Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, J.J., Shoichet, B.K. Relating protein pharmacology by ligand chemistry. Nature Biotech. 2007, 25, 197–206.
20. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. Basic local alignment search tool. J. Mol. Biol. 1990, 215, 403–10.
21. Hert, J., Keiser, M.J., Irwin, J.J., Oprea, T.I., Shoichet, B.K. Quantifying the relationships among drug classes. J. Chem. Inf. Model. 2008, 48, 755–65.
22. Nidhi, Glick, M., Davies, J.W., Jenkins, J.L. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. J. Chem. Inf. Model. 2006, 46, 1124–33.
23. Mestres, J., Martín-Couce, L., Gregori-Puigjané, E., Cases, M., Boyer, S. Ligand-based approach to in silico pharmacology: Nuclear receptor profiling. J. Chem. Inf. Model. 2006, 46, 2725–36.
24. Izrailev, S., Farnum, M.A. Enzyme classification by ligand binding. Proteins 2004, 57, 711–24.
25. Pruitt, K.D., Tatusova, T., Maglott, D.R. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007, 35, D61–5.
26. Higgins, D.G., Sharp, P.M. CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. Gene 1988, 73, 237–44.

27. Caffrey, D.R., Dana, P.H., Mathur, V., Ocano, M., Hong, E.J., Wang, Y.E., Somaroo, S., Caffrey, B.E., Potluri, S., Huang, E.S. PFAAT version 2.0: A tool for editing, annotating, and analyzing multiple sequence alignments. BMC Bioinformatics 2007, 8, 381.
28. Clamp, M., Cuff, J., Searle, S.M., Barton, G.J. The jalview java alignment editor. Bioinformatics 2004, 20, 426–7.
29. Tamura, K., Dudley, J., Nei, M., Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evolution 2007, 24, 1596–9.
30. Lee, M.R., Dominguez, C. MAP kinase p38 inhibitors: Clinical results and an intimate look at their interactions with p38alpha protein. Curr. Med. Chem. 2005, 12, 2979–94.
31. Vulpetti, A., Bosotti, R. Sequence and structural analysis of kinase ATP pocket residues. Farmaco. 2004, 59, 759–65.
32. Kothe, M., Kohls, D., Low, S., Coli, R., Rennie, G.R., Feru, F., Kuhn, C., Ding, Y.H. Selectivity-determining residues in Plk1. Chem. Biol. Drug Des. 2007, 70, 540–6.
33. Kothe, M., Kohls, D., Low, S., Coli, R., Cheng, A.C., Jacques, S.L., Johnson, T.L., Lewis, C., Loh, C., Nonomiya, J., Sheils, A.L., Verdries, K.A., Wynn, T.A., Kuhn, C., Ding, Y.H. Structure of the catalytic domain of human polo-like kinase 1. Biochemistry 2007, 46, 5960–71.
34. Sheinerman, F.B., Giraud, E., Laoui, A. High affinity targets of protein kinase inhibitors have similar residues at the positions energetically important for binding. J. Mol. Biol. 2005, 352, 1134–56.
35. Ortiz, A.R., Gomez-Puertas, P., Leo-Macias, A., Lopez-Romero, P., Lopez-Viñas, E., Morreale, A., Murcia, M., Wang, K. Computational approaches to model ligand selectivity in drug design. Curr. Top. Med. Chem. 2006, 6, 41–55.
36. Lichtarge, O., Bourne, H.R., Cohen, F.E. An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. 1996, 257, 342–58.
37. Armon, A., Graur, D., Ben-Tal, N. ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. J. Mol. Biol. 2001, 307, 447–63.
38. Schmitt, S., Kuhn, D., Klebe, G. A new method to detect related function among proteins independent of sequence and fold homology. J. Mol Biol. 2002, 323, 387–406.
39. Ferre, F., Ausiello, G., Zanzoni, A., Helmer-Citterich, M. Functional annotation by identification of local surface similarities: A novel tool for structural genomics. BMC Bioinformatics 2005, 6, 194.
40. Gold, N.D., Jackson, R.M. Fold independent structural comparisons of protein–ligand binding sites for exploring functional relationships. J. Mol. Biol. 2006, 355, 1112–24.
41. Kuhn, D., Weskamp, N., Hüllermeier, E., Klebe, G. Functional classification of protein kinase binding sites using Cavbase. Chem. Med. Chem. 2007, 2, 1432–47.
42. Gold, N.D., Deville, K., Jackson, R.M. New opportunities for protease ligand-binding site comparisons using SitesBase. Biochem. Soc. Trans. 2007, 35, 561–5.
43. Pastor, M., Cruciani, G. A novel strategy for improving ligand selectivity in receptor-based drug design. J. Med. Chem. 1995, 38, 4637–47.
44. Kastenholz, M.A., Pastor, M., Cruciani, G., Haaksma, E.E.J., Fox, T. GRID/CPCA: A new computational tool to design selective ligands. J. Med. Chem. 2000, 43, 3033–44.
45. Myshkin, E., Wang, B. Chemometrical classification of ephrin ligands and Eph kinases using GRID/CPCA approach. J. Chem. Inf. Comput. Sci. 2003, 43, 1004–10.
46. Vulpetti, A., Crivori, P., Cameron, A., Bertrand, J., Brasca, M.G., D'Alessio, R., Pevarello, P. Structure-based approaches to improve selectivity: CDK2-GSK3beta binding site analysis. J. Chem. Inf. Model. 2005, 45, 1282–90.
47. Yoon, S., Smellie, A., Hartsough, D., Filikov, A. Computational identification of proteins for selectivity assays. Proteins. 2005, 59, 434–43.
48. Sherman, W., Tidor, B. Novel method for probing the specificity binding profile of ligands: Applications to HIV protease. Chem. Biol. Drug Des. 2008, 71, 387–407.
49. Radhakrishnan, M.L., Tidor, B. Specificity in molecular design: A physical framework for probing the determinants of binding specificity and promiscuity in a biological environment. J. Phys. Chem. B 2007, 111, 13419–35.
50. Watkins, R.E., Wisely, G.B., Moore, L.B., Collins, J.L., Lambert, M.H., Williams, S.P., Willson, T.M., Kliewer, S.A., Redinbo, M.R. The human nuclear xenobiotic receptor PXR: Structural determinants of directed promiscuity. Science 2001, 292, 2329–33.

51. Jacobs, M.D., Caron, P.R., Hare, B.J. Classifying protein kinase structures guides use of ligand-selectivity profiles to predict inactive conformations: Structure of lck/imatinib complex. Proteins 2007, 70, 1451–60.
52. Cee, V.J., Cheng, A.C., Romero, K., Bellon, S., Mohr, C., Whittington, D.A., Bready, J., Caenepeel, S., Coxon, A., Deak, H.L., Hodous, B.L., Kim, J.L., Lin, J., Nguyen, H., Olivieri, P.R., Patel, V.F., Wang, L., Hughes, P., Geuns-Meyer, S., Pyridyl-pyrimidine benzimidazole derivatives as potent, selective, and orally bioavailable inhibitors of Tie-2 kinase, Bioorg. Med. Chem. Ltrs., in press.
53. Coleman, R.G., Salzberg, A.C., Cheng, A.C. Structure-based identification of small molecule binding sites using a free energy model. J. Chem. Inf. Model. 2006, 46, 2631–7.
54. Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Delivery Rev. 2001, 46, 3–26.
55. Palm, K., Stenberg, P., Luthman, K., Artursson, P. Polar molecular surface properties predict the intestinal absorption of drugs in humans. Pharm. Res. 1997, 14, 568–71.
56. 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, 2615–23.
57. Oprea, T.I., Davis, A.M., Teague, S.J., Leeson, P.D. Is there a difference between leads and drugs? A historical perspective. J. Chem. Inf. Comput. Sci. 2001, 41, 1308–15.
58. Lipinski, C. Filtering in drug discovery. Ann. Reports Comput. Chem. 2005, 1, 155–68.
59. Hajduk, P.J., Huth, J.R., Fesik, S.W. Druggability indices for protein targets derived from NMR-based screening data. J. Med. Chem. 2005, 48, 2518–25.
60. Hajduk, P.J., Huth, J.R., Tse, C. Predicting protein druggability. Drug Discov. Today 2005, 10, 1675–82.
61. Hopkins, A.L., Groom, C.R. The druggable genome. Nat. Rev. Drug Discov. 2002, 1, 727–30.
62. Overington, J.P., Al-Lazikani, B., Hopkins, A.L. How many drug targets are there? Nature Rev. Drug Discov. 2006, 5, 993–6.
63. Sakharkar, M.K., Sakharkar, K.R., Pervaiz, S. Druggability of human disease genes. Int. J. Biochem. Cell Biol. 2007, 39, 1156–64.
64. Hambly, K., Danzer, J., Muskal, S., Debe, D.A. Interrogating the druggable genome with structural informatics. Mol. Divers. 2006, 10, 273–81.
65. Russ, A.P., Lampel, S. The druggable genome: An update. Drug Discov. Today 2005, 10, 1607–10.
66. Orth, A.P., Batalov, S., Perrone, M., Chanda, S.K. The promise of genomics to identify novel therapeutic targets. Expert Opin. Ther. Targets 2004, 8, 587–96.
67. Kubinyi, H. Drug research: Myths, hype and reality. Nature Rev. Drug Disc. 2003, 2, 665–8.
68. Fauman, E.B., Hopkins, A.L., Groom, C.R. Structural bioinformatics in drug discovery. In: Weissig, H., Bourne, P., editors. Structural Bioinformatics. Hoboken, NJ: Wiley-Liss; 2003, p. 477–98.
69. Landon, M.R., Lancia Jr., D.R., Yu, J., Thiel, S.C., Vajda, S. Identification of hot spots within druggable binding regions by computational solvent mapping of proteins. J. Med. Chem. 2007, 50, 1231–40.
70. An, J., Totrov, M., Abagyan, R. Pocketome via comprehensive identification and classification of ligand binding envelopes. Mol. Cell Proteomics. 2005, 4, 752–61.
71. An, J., Totrov, M., Abagyan, R. Comprehensive identification of druggable protein ligand binding sites. Genome Inform. 2004, 15, 31–41.
72. Nayal, M., Honig, B. On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. Proteins 2006, 63, 892–906.
73. Pettit, F.K., Bare, E., Tsai, A., Bowie, J.U. HotPatch: A statistical approach to finding biologically relevant features on protein surfaces. J. Mol. Biol. 2007, 369, 863–79.
74. Hardy, J.A., Lam, J., Nguyen, J.T., O'Brien, T., Wells, J.A. Discovery of an allosteric site in the caspases. Proc. Natl. Acad. Sci. USA 2004, 101, 12461–6.
75. Joughin, B.A., Tidor, B., Yaffe, M.B. A computational method for the analysis and prediction of protein: Phosphopeptide-binding sites. Protein Sci. 2005, 14, 131–9.
76. Halgren, T. New method for fast and accurate binding-site identification and analysis. Chem. Biol. Drug Des. 2007, 69, 146–8.
77. Friesner, R.A., Murphy, R.B., Repasky, M.P., Frye, L.L., Greenwood, J.R., Halgren, T.A., Sanschagrin, P.C., Mainz, D.T. Extra precision Glide: Docking and scoring incorporating a model of hydrophobic enclosure for protein–ligand complexes. J. Med. Chem 2006, 49, 6177–96.

78. Cheng, A.C., Coleman, R.G., Smyth, K.T., Cao, Q., Soulard, P., Caffrey, D.R., Salzberg, A.C., Huang, E.S. Structure-based maximal affinity model predicts small-molecule druggability. Nature Biotech. 2007, 25, 71–5.
79. Davis, A.M., Teague, S.J. Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis. Agnew. Chem. Int. Ed. 1999, 38, 736–49.
80. Sharp, K.A., Nicholls, A., Fine, R.F., Honig, B. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. Science 1991, 252, 106–9.
81. Cheng, Y.-K., Rossky, P.J. Surface topography dependence of biomolecular hydrophobic hydration. Nature 1998, 392, 696–9.
82. Southall, N.T., Dill, K.A. The mechanism of hydrophobic solvation depends on solute radius. J. Phys. Chem. B 2000, 104, 1326–31.
83. De Young, L.R., Dill, K.A. Partitioning of nonpolar solutes into bilayers and amorphous *n*-alkanes. J. Phys. Chem. 1990, 94, 801–9.
84. Copeland, R.A., Pompliano, D.L., Meek, T.D. Drug-target residence time and its implications for lead optimization. Nature Rev. Drug Discov. 2006, 5, 730–9.
85. Brown, S.P., Hajduk, P.J. Effects of conformational dynamics on predicted protein druggability. Chem. Med. Chem. 2006, 1, 70–2.
86. Eyrisch, S., Helms, V. Transient pockets on protein surfaces involved in protein–protein interaction. J. Med. Chem. 2007, 50, 3457–64.
87. Braisted, A.C., Oslob, J.D., DeLano, W.L., Hyde, J., McDowell, R.S., Waal, N., Yu, C., Arkin, M.R., Raimundo, B.C. Discovery of a potent small molecule IL-2 inhibitor through fragment assembly. J. Am. Chem. Soc. 2003, 125, 3714–5.
88. Raimundo, B.C., Oslob, J.D., Braisted, A.C., Hyde, J., McDowell, R.S., Randal, M., Waal, N.D., Wilkinson, J., Yu, C.H., Arkin, M.R. Integrating fragment assembly and biophysical methods in the chemical advancement of small-molecule antagonists of IL-2: An approach for inhibiting protein-protein interactions. J. Med. Chem. 2004, 47, 3111–30.
89. Thanos, C.D., Randal, M., Wells, J.A. Potent small-molecule binding to a dynamic hot spot on IL-2. J. Am. Chem. Soc. 2003, 125, 15280–1.
90. Thanos, C.D., DeLano, W.L., Wells, J.A. Hot-spot mimicry of a cytokine receptor by a small molecule. Proc. Natl. Acad. Sci. USA 2006, 103, 15422–7.
91. Joseph-McCarthy, D. Structure-based lead optimization. Ann. Reports Comp. Chem. 2005, 1, 169–83.
92. Armstrong, K.A., Tidor, B., Cheng, A.C. Optimal charges in lead progression: A structure-based neuraminidase case study. J. Med. Chem. 2006, 49, 2470–7.
93. Hardy, J.A., Wells, J.A. Searching for new allosteric sites in enzymes. Curr. Opin. Struct. Biol. 2004, 14, 706–15.