3862 J. Med. Chem. 2010, 53, 3862–3886
DOI: 10.1021/jm900818s

Journal of Medicinal Chemistry Perspective

Molecular Shape and Medicinal Chemistry: A Perspective

Anthony Nicholls,*[,†] Georgia B. McGaughey,[‡] Robert P. Sheridan,[‡] Andrew C. Good,[§] Gregory Warren,[†] Magali Mathieu,[∥] Steven W. Muchmore,[⊥] Scott P. Brown,[⊥] J. Andrew Grant,[#] James A. Haigh,[†] Neysa Nevins,[∞] Ajay N. Jain,[×] and Brian Kelley[†]

[†]*OpenEye Scientific Software, Inc., 9d Bisbee Court, Santa Fe, New Mexico 87508,* [‡]*Merck & Co. Inc., Rahway, New Jersey,* [§]*Genzyme Corp., Cambridge, Massachusetts,* [∥]*Sanofi-Aventis, Vitry, France,* [⊥]*Abbott Laboratories, Abbott Park, Illinois,* [#]*AstraZeneca, Mereside, U.K.,* [∞]*GlaxoSmithKline, King of Prussia, Pennsylvania, and* [×]*University of California, San Francisco, California*

Received June 5, 2009

Shape and Medicinal Chemistry

In his philosophic musings "Meditations" the Emperor Marcus Aurelius asks "This thing, what is it in itself, in its own constitution? What is its substance and material? And what its causal nature?"[1] The history of chemistry, and in particular medicinal chemistry, is an elaboration of these three questions as applied to molecules: "What is the essence of a molecule? What is it made of? What will it do?"

In trying to answer these questions, and thereby describe molecules, we create languages. Primo Levi, the great writer and chemist, complained in 1984 that there were only three accepted ways to describe a molecule and none of them were very good: the ambiguous molecular formula, the nonlexical chemical graph, and the (often obscure) chemical name.[2] Yet, because these are the ways we describe a molecule's "constitution", these dominate our approaches to predicting what a molecule will do. Even SMILES,[3] developed by David Weininger shortly after Levi's lament, and intended to be a real lexicographic description, only facilitated methods that rely on the counting of elements of composition, e.g., chemical rules of thumb, classification algorithms, druglike filters (e.g., the ubiquitous rule of five[4]), 2D QSAR, or molecular fingerprints. While we may have elaborated beyond the elemental to include graph-related properties (e.g., aromaticity, hydrophobicity, hydrophilicity, hydrogen bond donors and acceptors, and so forth), these are seldom fundamental and often just opinions on how molecules behave.

To further our ability to predict, we have to consider other "essential" aspects of a molecule, in particular its three-dimensional form. It is a subject of continuing investigation as to how best to capture this "essence", and this Perspective details the contribution of molecular shape. Shape is not the only approach; for instance, the well-known concept of 3D pharmacophores has proved very successful.[5] Yet pharmacophores describe atoms or sets of atoms as points in space, and molecules are more than that; they are volumes and surfaces. Approaches that focus on shape, as described here, go beyond pharmacophoric methods in both utility and generality. And while some have tried to use pharmacophores to describe shape,[6] such efforts have not been very successful; shape is simply a different descriptive paradigm.

So what do we really mean by shape? There is a simple, universal meaning to the concept as the coincidence of volumes (Figure 1) that can also be extended to surfaces. Despite this precise and very general definition, there are many less general and more limited interpretations. We have avoided considering these approaches in order to present a more cohesive perspective, although there are excellent reviews on these various methods.[7] We do, however, include an analysis of attempts to approximate shape. Such methods are inevitably "lossy"; i.e., they trade information for the expediency of computational simplicity and speed. Any attempt to answer the first of Aurelius' questions is always going to be incomplete; as Kuhn points out, there are always new levels of understanding in science.[8] Yet finding a good and useful essence is hard work, and so we consider if these approximate methods are worth the loss of verisimilitude.

Initially the motivation for shape in drug discovery was virtual screening; if two molecules have a similar shape, perhaps they have similar properties. Despite Quine's adage that "exploiting the similarity concept is a sign of immature science",[9] shape similarity is now quite a mature approach. Yet the truest measure of an idea is not only its usefulness as originally conceived but also how its ambit expands over time, something this article attempts to chronicle. In addition to lead discovery, we have asked developers of theory and practitioners of methods to describe the application of molecular shape in areas as diverse as crystallographic refinement, docking and pose prediction, clustering, library design, and lead optimization. Finally, we ask what the new directions for shape in molecular modeling might be. Does shape provide a viable new language for chemistry, or is that still out of reach? Clearly this is worth a meditation.

Shape and Virtual Screening

The term "virtual screening" is fairly new. A SciFinder search suggests the first appearance of this phrase was in the 1990's,[10] but the idea has been around for a long time. The concept of using 3D similarity (sometimes using shape alone, sometimes using atom typing, i.e., assignment of chemical character to an atom or group of atoms or the fields emanating from molecules) as a basis for virtual screening is integral to a computational chemist's tool chest. As Clark suggested,[10] companies that were pioneers in this area have many success stories. Indeed, Merck Research Laboratories (MRL) has been developing virtual screening methods for decades[11−13] and has several published examples where 3D similarity has been applied in virtual screening and projects have been

*To whom correspondence should be addressed. Phone: 505-473-7385, extension 61. Fax: 505-473-0833. E-mail: anthony@eyesopen.com.

 Published on Web 02/16/2010

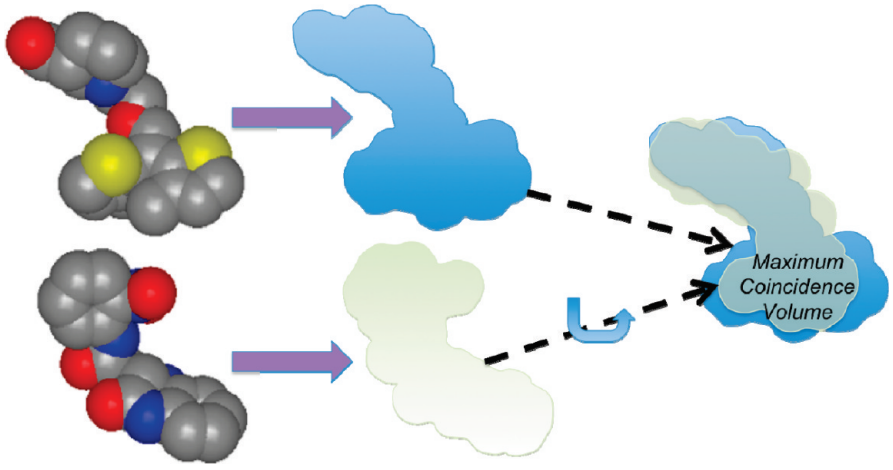

Figure 1. Illustration of a fundamental definition of shape similar, derived from the alignment that achieves an optimal overlap of objects. The mismatch volume between two objects is a true mathematical metric distance, i.e., obeys the triangle inequality that says the distance from object A to object C cannot be greater than the distance from A to B plus B to C nor less than the difference between these distances. However, the optimal overlap leads the more intuitive Shape Tanimoto (ST), i.e., the ratio of the overlap to the absolute difference of the sum of the self-overlaps and optimal overlap. It has the useful character of ranging from 1.0 (perfect overlap) to 0.0 (no overlap).

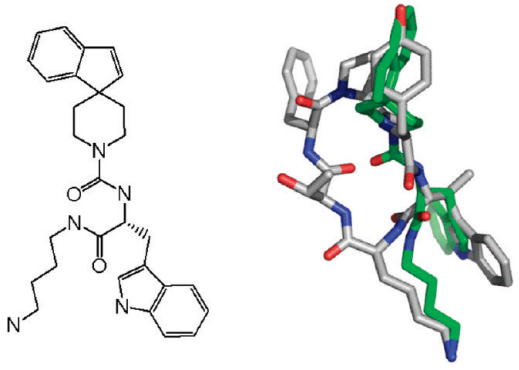

Figure 2. Compound **1** (left-hand image) and superposition of compound **1** (green) with probe ligand (white) using SQW.[13]

thereby advanced.[14−16] Most of these early efforts helped bridge the gap from a peptide-like lead to a drug-like lead.

MRL's first published application of virtual screening was in the non-peptide fibrinogen receptor antagonist program.[14] Starting from the endogenous Arg-Gly-Asp motif, a virtual screen of the corporate database identified many non-peptides that mimicked this group. Some of these were tested, and one turned out to be a 27 μM (IC_{50}) lead. In another example, a query was constructed from key amino acids of somatotropin release-inhibitor factor (SRIF) plus additional "sphere points" that defined salient electrostatic and volume regions. A virtual screen of Merck's flexibase[17] of over 1 million compounds using this query identified compound **1** (Figure 2), which was found to have measurable activity.[16,18] This compound was ranked 41 out of >1 million compounds by SQW (SemiQuantitative reWrite).[13] SQW is the second generation of a proprietary 3D similarity/superposition program written in-house at MRL. The original program SQ (for SemiQuantitative) was written in the late 1990's. SQ/SQW operates on a rigid molecules represented as heavy atoms that have been classified into seven physiochemical types (cation, anion, etc.). First, a clique matching algorithm is used to generate many orientations of a candidate molecule onto a target molecule. Second, a Nelder−Mead simplex algorithm adjusts the orientation of the candidate molecule to optimize the score. An analogue of compound **1** was the first potent and selective small molecule somatostatin receptor 2 (SST2) agonist reported at the time. The superposition of compound **1** (green) with the probe ligand (white) is shown in Figure 2.

MRL was not the only group to develop shape algorithms for the primary application in lead identification, but we did develop one of the earliest 3D superposition methods, called SEAL,[11] which took into account charges and volumes. There are numerous independent software vendors (ISVs) and academic groups that have subsequently published in this area, and we refer you to the following references for an overview.[10,19] In 2004 we undertook a large scale comparison of purchasable shape-based methods for use in virtual screening.[10,20] In retrospect, we did find, as suggested by others in the literature, that there are many pitfalls and "gotchas" connected with the whole enterprise of method comparison that make it hard to arrive at robust conclusions. We refer interested readers to a special *Journal of Computer-Aided Molecular Design* issue "Evaluation of Computational Methods: Insights, Philosophies and Recommendations"[21] for many suggestions on how to properly conduct an evaluation.

The most important conclusion from our study is that, within the limits of retrospective screening, knowing the structure of an active ligand is better than knowing the atomic structure of its receptor. This is true if what one cares about is how many actives are retrieved and does not, for instance, need to find a plausible docking mode. We are not the first to say this; our conclusions are in agreement with earlier findings on this topic.[22] It seems to be generally true regardless of which database one screens[20,23] or which ligand or protein structure one uses for the virtual screen.[24] As time went on we realized, based on valid critiques of our study, that we needed to change the way we compared methods, the most important of which had to do with the set of targets we used: (1) we sought to have enough targets to minimize the uncertainties due to the composition of the target set and (2) we would have to choose only those targets where the number of actives was fairly large. Simulation studies in MRL by Truchon and Bayly[25] also reinforced the need for more actives. We therefore developed a set of 47 small molecule targets such that the number of diverse actives in the MDDR was >20. The

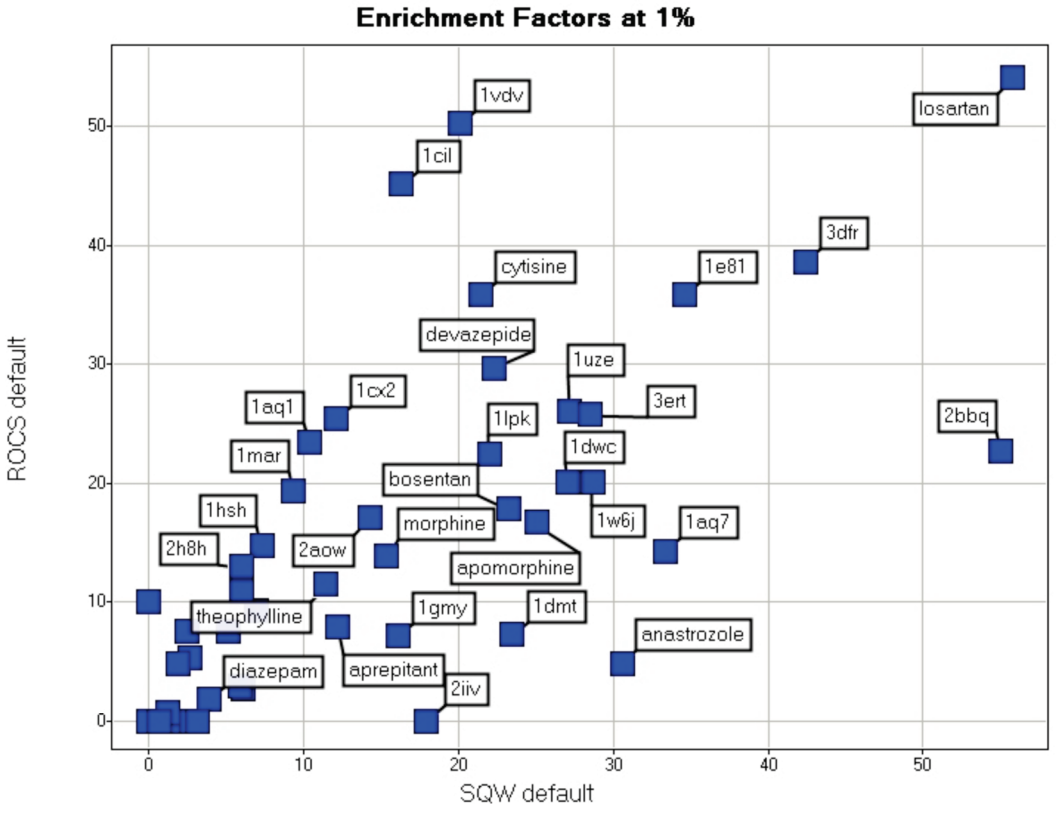

Figure 3. Enrichment factors (EF) at 1% for ROCS and SQW results for over 47 unique targets. Enrichment here is the ratio of the number of actives at a given percent of the database to the expected number of such.

majority of the targets have cocrystallized ligands in the Protein Data Bank (PDB),[26] but some are derived 3D geometries using CORINA.

Given this new set, we carried out a number of studies[27] comparing various 2D and 3D similarity methods as virtual screening engines. Since it is apropos for this venue, we will focus on ROCS (rapid overlay of chemical structures) and SQW, which are both 3D similarity methods. In our hands 3D similarity methods seem to embody the best combination of finding the most actives in virtual screening and having those actives be diverse.[20] ROCS is considered a state of the art 3D similarity method. ROCS searches for optimal shape overlays, as illustrated in Figure 1. It uses atom-centered Gaussians to accurately represent volumes because such functions are much smoother than discrete "inside/outside" representations, e.g., molecules as fused spheres. As a consequence, the number of overlap maxima is much reduced, enabling approximations to the global maxima to be found quickly. It also includes the facility to match chemical types by representing atoms or groups of atoms as Gaussians of a given "type" or "color", for instance, rings, hydrogen bond donors and acceptors. It has a lot in common with SQW in the core concepts (atoms typed as hydrogen-bond donor, acceptor, etc.; atoms represented as Gaussian functions, i.e., a soft, extended function, rather than a one or zero function corresponding to a hard sphere), but the detailed implementation is different. One interesting addition in ROCS is the inclusion of a "ring" term, where molecular superpositions get extra credit if ring centroids are superimposed, regardless of the type of ring. Our findings are that while SQW and ROCS do not perform the same on any given target, the average performance over 47 targets is surprisingly similar (Figure 3).

Despite the fact that 3D similarity methods perform very well, by no means are we saying they are a panacea. We often state in regard to virtual screening methods that "everything works on something; nothing works on everything".[28] If one is in the lead finding stage of a program, 3D similarity may be the most straightforward method to obtain diverse leads.[29−31] However, if your protein target can adopt multiple conformations (because of inherent flexibility), one may be less successful retrieving a novel, active ligand using 3D similarity methods. This is the case for β-secretase (BACE), which is implicated in Alzheimer's disease. For instance, if you used the hydroxyethylamine ligand from the PDB code 2B8L[32] as a probe for virtual screening (which interacts directly with both catalytic aspartic acids Asp^{32} and Asp^{228} and occupies regions P1, P2, P3, and P1′), it would be nearly impossible to identify in the top rankings a spiropiperidine (PDB code 3FKT[33]) which interacts with the catalytic aspartic acids via a water molecule and does not occupy regions P2 or P3 at all. Furthermore, the best ROCS superposition of these two ligands (Figure 4a) does not even qualitatively overlay the ligands in the way one observes crystallographically (Figure 4b). Why is this? Clearly 3D similarity methods do not take into account the influence of the receptor and, by design, will maximize volume overlap between two molecules. This is in contrast to how the BACE ligands are bound in their cognate sites (Figure 4b). In the BACE example, they occupy different spatial regions of the enzyme. This is clearly a failure of the assumptions behind 3D similarity methods, and protein flexibility obviously makes this issue more severe. That said, one should not abandon the utility of 3D shape based similarity in such cases; rather, such a virtual screen should be complemented with additional computational methods such as docking.

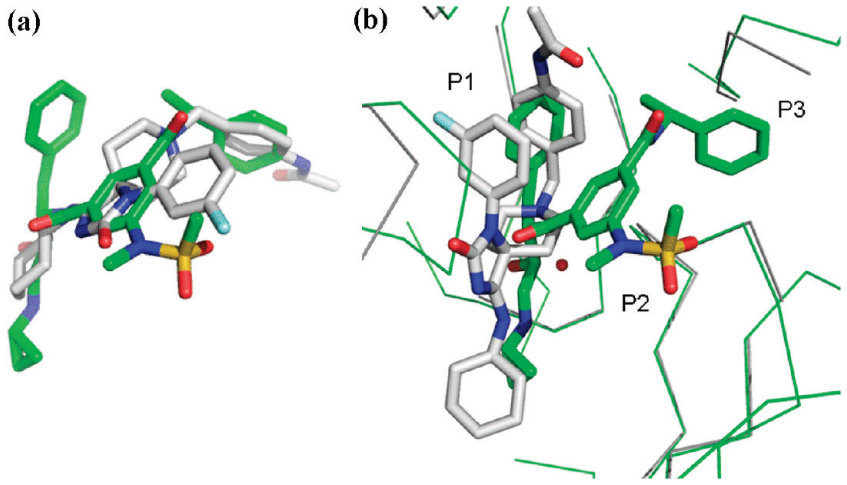

Figure 4. (a) Highest ranked ROCS superposition by Combo score (sum of the ST and raw color overlap) of ligands from PDB codes 3FKT (white) and 2B8L (green). (b) Aligned enzyme structures (3FKT and 2B8L) with cognate ligands. The key piperidine−water mediated hydrogen-bonding interaction with the catalytic acids is depicted as a red sphere.

In contrast to the failure exemplified in the BACE example, there are numerous examples where the shape of the ligand coupled with electrostatics to capture the polarity of the atoms has identified novel, noncongeneric hits via virtual screening.[14−16,29−31] Are the hits retrieved found for the "right reason"? For instance, if overlaid in the active site of the protein or enzyme, would the correct chemical features align in the active site in contrast to the BACE example depicted in Figure 4a? In the case where ROCS was used to find inhibitors of ZipA-FtsZ,[31] the authors subsequently crystallized one of the hits in ZipA and determined that ROCS predicted the binding mode. Induced fit, manifested as enzyme flexibility, was not integral to this binding motif and demonstrated that the hit retrieved in that study was found for the "right reason". That is, when overlaid in the active site of the protein, the correct chemical features align in the active site.

The emergence of public data sets such as DUD allows us to ask if there are trends in the proficiency of shape tools; for instance, do they work worse with active compounds that are flexible or better for active sites that are small? Preliminary evidence is that shape is fairly robust with respect to operational parameters. However, some trends can be observed. Figures 5 shows the performance of ROCS, measured by area under the Receiver Operator Characteristic (ROC) curve (AUC), over the cocrystal structures in DUD. Each symbol represents a query molecule in DUD, with its own particular set of decoys. We specifically asked the question how (A) heavy atom count, (B) the ratio of ROCS' so-called "color" atoms to heavy atoms, and (C) intrinsic ligand affinity affect the performance of this shaped-based method to retrieve active compounds. A number of other properties, such as charge, ligand flexibility, number of color atoms, and a measure of site polarity, were also tested and found to have little or no correlation with performance. The decreasing performance with respect to number of non-hydrogen atoms in the query seemed reasonable; i.e., perhaps the conformational space of larger molecules was harder to search. On the other hand there seems to be no correlation of AUC with the number of rotatable bonds in the query. Investigation of the ratio of color atoms to heavy atom count in the query suggests that query molecules that do not have enough color points tend to have poorer performance, which is consistent with observations that ROCS "shape" tends to be poorer than ROCS with color. Finally, there does not seem to be any significant relationship between the potency of the query on its particular target with the ability of ROCS to select more actives with the same activity. The results in Figure 5 are by no means final, as we have not formulated a data set where each set of probe ligands are of similar ligand affinity or molecular weight, etc.; however, a study focused on understanding those physicochemical properties is warranted. We have merely scratched the surface in exposing that some of these features may affect performance.

In conclusion, the application of 3D virtual screening methods has resulted in the identification of many active compounds in drug discovery programs. We and others have repeatedly demonstrated that 3D similarity-based virtual screening maintains enrichment in actives and increases the diversity in the compounds discovered. As a consequence, shape is integral to any drug discovery program. Finally, shape is a necessary requirement to facilitate the identification of both active and novel ligands in drug discovery, but it is not ultimately sufficient in the entire life cycle of a program where other factors, e.g., pharmacokinetic properties, may become more important and where shape has yet to play a significant role. In addition, there are clearly situations where shape is less useful. For instance, flexibility of the target enzyme or receptor can reduce the effectiveness of a 3D shape-based virtual screen, i.e., if there is less shape "coherency" between active molecules. In general practice, however, knowledge of merely one active compound can often outweigh the presumed advantages of an existing protein structure. Applications to molecules of a more fragment-like nature, where shape discrimination is subtler, are considered later in this Perspective in the context of library design.

Lead Optimization

The synthesis of compounds necessary for a detailed search of chemical space for lead optimization can make it much more resource intensive than lead discovery. In order to justify this cost, the yield, in terms of active compounds with improved physical properties, must be significantly higher than screening or virtual screening. This leaves chemists with a difficult balancing act; in order to modify the properties of the lead, they must change the molecule significantly. Yet to maintain activity, they must be conservative. For many years, chemists have been doing this through the knowledge of

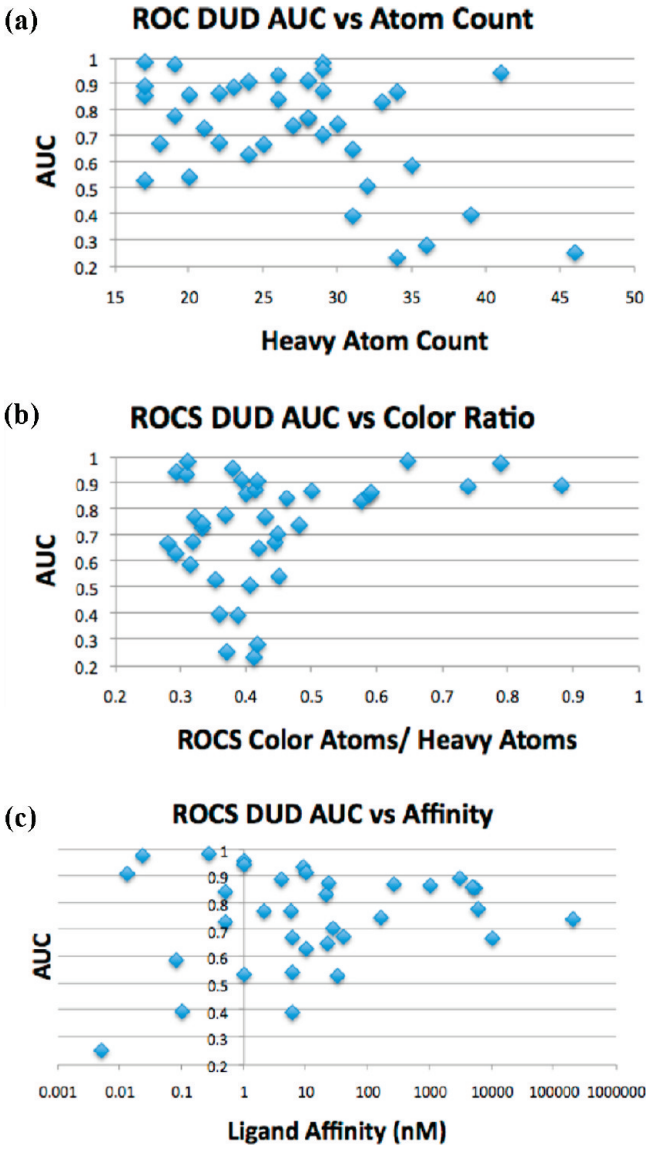

Figure 5. (a−c) Performance of ROCS over the DUD data set. DUD consists of a set of 40 protein−ligand systems, 38 of which has crystallographic coordinates. The data presented here are the results from taking the crystallographic ligand as the ROCS query in a virtual screening experiment against all other active ligands and the standard DUD decoys for that system. Presented are the AUC values, i.e., the area under the curve for each ROC curve, versus three ligand properties, namely, number of non-hydrogen atoms (a), ratio of the number of color points to this number (b), and the ligand affinity (c). Note the number of color points is the number of color Gaussians added to the shape optimization and depends on the number of chemical features, e.g., donors, acceptors, etc.

bioisosteres and other intuitively semiconservative fragment replacements. Shape-based fragment-similarity tools automate and extend this common practice. Rather than sampling the thousands of fragments that might come into a chemist's mind, such tools can search millions of fragments and winnow the list, so a modeler or chemist can assess the viability of just a few compounds, knowing that all of the fragment replacements have the same shape as the original fragment, balancing the need for change and the need for stability. At the same time, other physical properties, such as those that make up Lipinski's "rule of five", can be easily calculated as part of the lead optimization objective function. More so even than in lead discovery, shape can play an essential role in guiding project goals.

Central to such approaches has been the deconstruction of molecular databases using pseudoretrosynthesis.[34] The resulting fragment databases can then be used in bioisosteric de novo design through comparison of parts of the active lead with fragments abstracted from pharmacologically active molecules. The advantages of experimental fragment-based screening based on the lowering of molecular complexity have been extensively discussed.[34−36] Many parallels can be drawn to this from the perspective of fragment-based alignment. Fragments, being smaller, tend to be far more rigid in nature, rendering active conformation elucidation less of a challenge. Superposition is also dramatically simplified, since in addition to fewer degrees of freedom the fragmentation points provide strong bond alignment constraints. Further, users have the ability to focus on portions of active ligands where structure−activity relationships are more readily understood, again maximizing the information content of the resulting screen.

This fragment replacement approach has been applied using a variety of molecular similarity descriptors,[37−45] with the BROOD program pioneering the technique from the perspective of shape-based template comparison.[41] KIN[46,47] is a shape-based screening tool developed within Bristol-Myers Squibb (BMS) and will be discussed here in some detail. KIN's molecular similarity calculations are made using Gaussian-derived comparisons of electrostatic potential and shape as developed within the Richards group.[47−49] These scoring functions have then been incorporated into the DOCK docking program to exploit its flexible clique search framework.[50] For de novo bioisosteric replacements additional Gaussian constraints have been added to force superposition of the linker bonds that map the fragment to its parent molecule. In addition, exclusion Gaussians can be integrated to reflect regions of protein bulk, derived either explicitly from protein structure or implicitly from SAR. Each element of the scoring function can be weighted differently, and shape Gaussians can be "colored", i.e., assigned chemical character, to force the mapping together of critical binding functionalities. Table 1 is provided to illustrate how the atom-by-atom descriptions can end up looking in a KIN template file.

The tight geometric constraints intrinsic to bioisosteric clique searches allow for extremely rapid searches, ranging from 10^5 to 10^7 conformations per CPU hour depending on query complexity. Some results of this approach are shown in Figure 6, with three search examples using different facets of KIN's search capabilities, namely, carbamate, biphenyl, and benzamidine replacement in factor VIIa serine protease (PDB entry 2bz6). The result is a program able to provide a rapid turnaround of scaffold hopping ideas, readily understood by chemists, in the context of existing chemotypes and SAR. KIN has been used extensively as a bioisostere replacement search tool within BMS.

There remain many open questions as to whether shape can contribute in other ways to lead optimization, for example, to the calculation of observable trends within series, such as activity. Typical approaches to these difficult problems involve either first-principles physics approaches, scoring functions, or QSAR methodology, none of which have proven generally reliable. Given the importance of shape and electrostatics in lead optimization and because of their role in protein−ligand interaction, it seems reasonable to hope for advances beyond the current methods of bioisosteric replacement.

Table 1. Extract of the KIN Template File for Benzamidine Replacement (Figure 5c)[a]

Atom Number/Name		Coordinates			Type	Critical Region	Charge/Weight	KIN atom type
1	C1	11.9182	38.8306	31.3299	C.ar	2	0.0	mapping_atom
2	C2	11.5122	39.0316	30.0079	C.ar	5	0.0	linker
3	C3	12.2792	38.5546	28.9599	C.ar	2	0.0	mapping_atom
4	C4	13.1352	38.1706	31.5689	C.ar	2	0.0	mapping_atom
5	C5	13.9242	37.6886	30.5229	C.ar	2	0.0	mapping_atom
6	C6	13.4932	37.8946	29.2049	C.ar	2	0.0	mapping_atom
7	C7	14.2812	37.4046	28.0309	C.cat	2	0.0	mapping_atom
8	N1	15.4680	36.7800	28.1850	N.pl3	2	0.0	mapping_atom
9	N2	13.7642	37.6086	26.8089	N.pl3	4	10.0	donor
10	H1	11.3002	39.1802	32.1437	H	1	0.0	null_ligand
11	H2	11.9453	38.6887	27.9416	H	1	0.0	null_ligand
12	H3	13.4811	38.0256	32.5816	H	1	0.0	null_ligand
13	H4	14.8463	37.1688	30.7372	H	1	0.0	null_ligand
14	H5	16.0244	36.4308	27.3027	H	1	0.0	null_ligand
15	H6	14.2642	37.2933	25.9899	H	1	0.0	null_ligand
16	H7	12.8752	38.0774	26.7085	H	1	0.0	null_ligand
17	Si1	9.9280	39.9476	29.6480	Si	3	0.0	r_group
18	H8	9.7265	40.0254	28.1934	H	1	0.0	null_ligand
19	H9	8.7986	39.2335	30.2618	H	1	0.0	null_ligand
20	H10	10.0099	41.3061	30.2049	H	1	0.0	null_ligand
21	H11	15.8450	36.6333	29.1103	H	1	0.0	null_ligand
1	CB	15.1960	33.6780	24.4740	C.3	1	-10.0	null_ligand

[a] Atom lines have been color-coded to match the colors shown in the template figure. These files are extensions of the mol2 file format. Atoms numbers, names, coordinates, and types are unchanged from the original format conventions. The critical region field is used to define atoms/clusters of atoms that must be present in any given target clique match prior to superposition onto the template. Region 1 is used to place atoms that will be ignored for clique mapping. For this search the rgroup atom (region 3), linker atom (region 5), a designated benzamidine amino atom (region 4), and one of the remaining heavy atoms (region 2) must be mapped simultaneously for successful clique extraction. The charge field is used as the color-weighting field when chemical matching is used in place of electrostatic similarity. It can also be used to weight exclusion regions by setting a negative weight to less than −10.0. For this search no electrostatics was used, so this field has been zeroed apart from the donor atom-matching portion and the single exclusion atom shown (the full template file contains 40 exclusion atoms). Note that no weighting has been applied for group and linker atoms, as they form a separate term in the overall KIN similarity (linker bond similarity) and are thus weighted independently in the KIN calculation input file (along with weighting terms for shape and electrostatic/color mapping similarity). The final field is the KIN atom type. These atom types are defined in an extensively modified version of DOCK's chemical definitions (chem.defn) file. For this search rgroup (atom that marks the disconnection point of the fragment, always Si in KIN), linker (atom linking rgroup connection point), donor (hydrogen bond donor), mapping atoms (atoms that must be mapped geometrically without reference to target atoms type), and null_ligand (atoms to ignore) have been used.

Protein Crystallography

Crystallographers have always worked with shapes. They look at an electron density map, cut or contoured at a given level, and interpret the shape to position the residue atoms that comprise the protein or nucleic acid being studied. Examples of tools that help to correctly place atoms, in particular residue atoms, are such well-known graphic programs as O,[51] Coot,[52] and Quanta.[53] These graphics packages come with options to help choose the correct rotamer and ensure that the geometry of each residue remains inside reasonable boundaries. Further refinement between the structure factors of the atoms and the measured data is then required before another round of visual checking of the agreement between density and model can occur. And yet because the "shapes" are familiar, i.e., from a small repertoire of possible amino acid shapes, few crystallographers would make the claim or even realize they use shape.

Difficulties arise when placing small molecule ligand atoms in density. Protein crystals rarely diffract to a high enough atomic resolution to precisely locate all atoms. The ligands themselves are often not clearly and completely defined in their binding site, a condition usually attributed to low occupancy or high mobility. Consequently, protein crystallographers are often faced with a vague shape, roughly the size of the ligand, in which to place a molecule with which they are not familiar. In addition, chemists usually describe ligands with one- or two-dimensional formats, while crystallographic programs typically only understand three-dimensional (3D) formats, in particular the PDB format. Protein refinement programs require parameter and topology information for the ligand in order to handle ligands appropriately. Most programs try to guess this information from a coordinate file usually with poor results, especially since the typical resolution of protein crystallographic data precludes the inclusion of hydrogen atoms. So to solve a ligand structure, the crystallographer needs to obtain a 3D structure of the ligand, determine relevant planarity and torsion information, and use graphical tools that allow for the fine-tuning of a ligand conformation in the visible density. This is a tricky problem and sometimes results in a "nonideal" conformation when using weak density. This problem is easily compounded by stereochemical ambiguities, saturated cycles, or worse, cyclic ligands.

Various attempts have been made to overcome some or all of these hurdles, most of them implicitly using the shape of the remaining electron density when the protein has been resolved. The Quanta application X-Ligand[54] first centers the ligand on an unoccupied patch of electron density and then searches the ligand conformational space to find the best fit to density. The graphical program COOT also searches clusters of density not occupied by the protein and tries to fit the ligand by comparing rigid body refinement values at each of the potential sites.[52] If a ligand dictionary has been provided, trial conformations are generated using the rotatable bonds to enhance the fit between

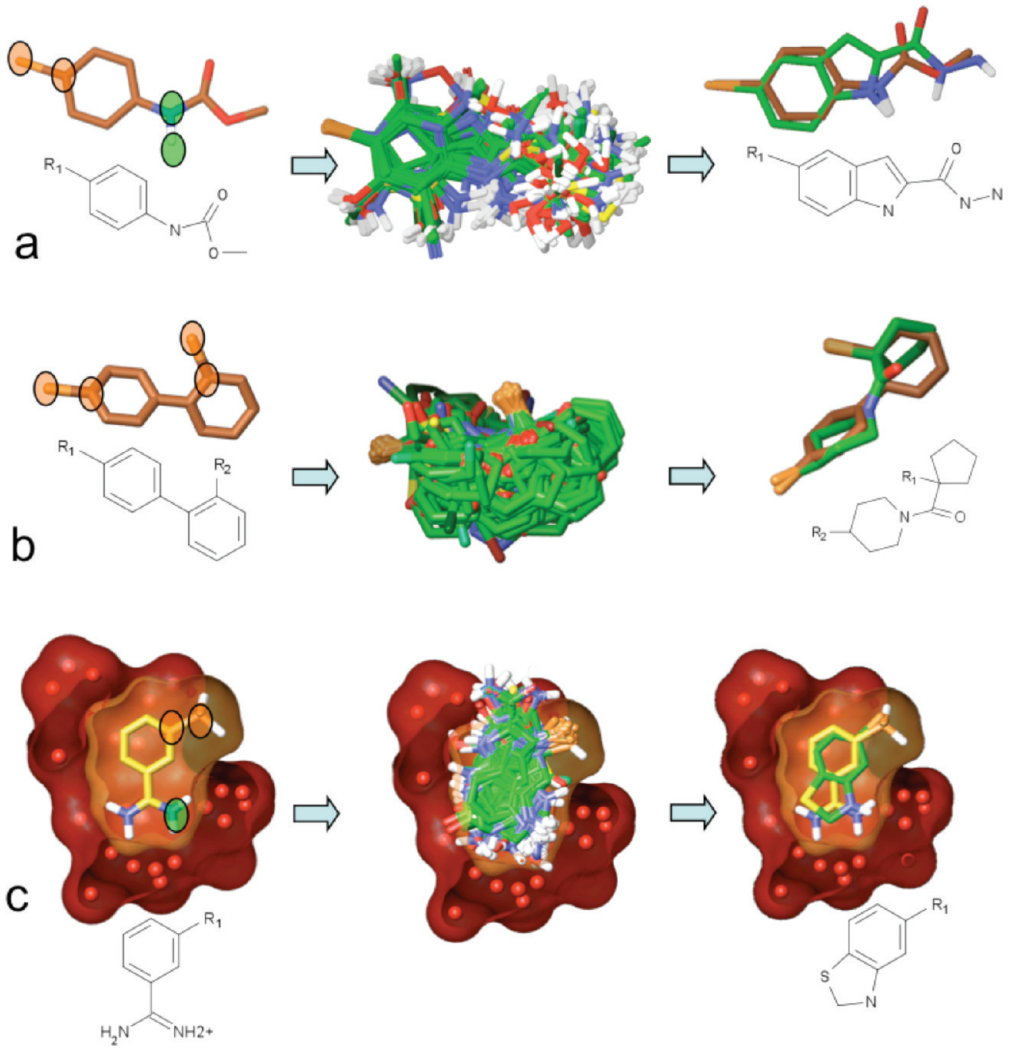

Figure 6. (a−c) Three search examples using KIN. Each example shows the template molecule with constraints applied (orange circles highlight linker Gaussians; green circles highlight critical pharmacophoric "colored" Gaussian constraints), top 100 hits superimposed (shown with green carbons), and one sample hit superimposed onto the template. The first linker bond database searched here contains $\sim 8 \times 10^6$ fragment conformers, the second linker database $\sim 10^7$. All timings were taken on a single AMD Opteron 2.1 GHz CPU. (a) Carbamate replacement search. All charges in the template are turned off except carbamate NH hydrogen, and chemical definitions were set so hydrogens attached to N/O are defined as special. Donor N are also set to be critical donor. Search time was 4 h. (b) Caveat[108] style search on biphenyl core template. Loose SAR is assumed in core region (low shape (0.33) and electrostatic potential (0.0) Gaussian score weighting assigned, while both linker bond Gaussian weightings are left at 1.0). Search time was 1.5 h. (c) Benzamidine replacement search. Search template was abstracted from the PDB structure 2bz6 factor VIIa serine protease structure.[109] Protein heavy atoms within 4 Å of the template were used as an exclusion Gaussian region. Benzamidine NH2 mapping carboxylate within S1 subsite was defined as a critical donor, with only neutral replacement fragments permitted. Linker Gaussian weighting was lowered (0.5) to allow more variability in linker bond vector position, and the linking vector was switched to the meta position of the benzamidine phenyl group. Chemical match scoring was used, not electrostatics, with chemical definitions modified to allow halide/sulfur donors. Search time was ~7 h.

ligand and potential site. The methods used by these programs do not easily handle different stereochemistries or configurations, as all variants have to be explicitly defined, built, and placed in density.

More recently, both GlobalPhasing, which uses a molecular replacement approach that works best on rigid ligands, and the PrimeX method[55] from Schrödinger, which uses a docking approach,[56] have entered the field. Unfortunately neither the manual fitting method nor automated methods (with the possible exception of PrimeX) monitor geometric strain of the ligand during fitting. This lack of monitoring can result in "nonideal" conformations that appear to fit the density (or shape) but have an unrealistically strained geometry, primarily because crystallographers lack the utility of amino acid Ramachandran maps. Can a shape-based method, complete with strain analysis, do better? There is substantial evidence that it can, which we will describe here in two ways: first with an example of how the shape-based program AFITT[57] straightforwardly produces low-strain conformations in a prospective project and, second, with a retrospective analysis of its ability to find lower energy conformations of ligands from protein−ligand complexes in the PDB.

As illustrated in Figure 7, AFITT uses shape for three of the tasks required for fitting a small molecule ligand: (1) identification of the density belonging to the small molecule, (2) fitting low-strain 3D conformations of the molecule to the density, and (3) optimizing a 3D conformation of the molecule to maximize the shape overlap between the molecular shape and the electron density while preventing a large increase in geometric strain.

For the first stage a "search density" is constructed by subtracting it from the resolved portion of the structure. The remaining density is queried to find pieces that have a molecular volume similar to that of the ligand, as determined by either estimating the volume given 1D or 2D input or calculated using the Gaussian method of Grant and Pickup[58] for 3D input. Regions with higher density (σ) levels have a higher probability of containing atoms. As a result, ligand density detection can be automated by parsing σ levels starting from a high value downward until a volume of density matching the volume of the ligand is discovered that is close to the protein surface. The major difference between this approach and the methods described previously is the identification of regions by isocontour surface, the AFITT technical term being a "blob". Contained within the blob is a piece of density that can be manipulated as an abstract object; e.g., it can be merged with other blobs when density for the ligand is fragmented or edited/sculpted in cases where a low σ level was chosen and water or noise density is now associated with the blob. Once a final collection of blobs is chosen, the shape of the contained electron density is passed to the second stage.

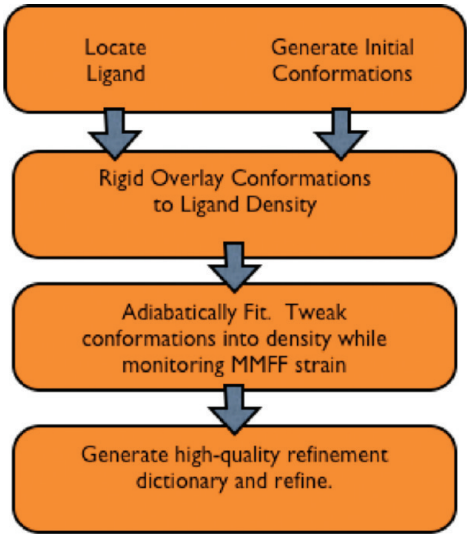

Figure 7. Workflow for AFITT ligand fitting.

During the second stage a functional representation (Gaussian) of the electron density shape is created and used as the query for a rigid shape-based superposition between the represented density and an extensive, if not exhaustive, ensemble of low energy ligand conformations. The particulars of this are the same as for the program ROCS[59] and result in a very rapid evaluation even over thousands of typically low energy conformations, taking a few seconds at most. Superpositions for each rigid conformation are ranked by overlap, and the best matches are passed to the optimization stage.

The last stage optimizes and relaxes the rigid superposition of the ligand conformation to the selected electron density. It does this under the combined influence of the Merck molecular force field and a gradually increasing component of the shape of the electron density. AFITT avoids overstraining the ligand by continually checking the induced strain versus the gain in shape matching, ensuring a limited geometric strain while maximizing the shape complementarity between ligand and electron density.[60] This fitting is referred to as an "adiabatic" fit because slowly changing the mixing of strain and shape prevents "jumping" to a different conformation (energy well), something that otherwise can occur when a fixed ratio of the two forces is applied (see Figure 8).

Prospective Case. As noted above, human β-secretase (BACE) is studied in the context of Alzheimer's disease because it cleaves amyloid precursor protein to generate Aβ peptide. Inhibitors usually bind above the catalytic aspartates, that are situated at the bottom of a binding site cleft. A "flap", in an open position in the apo structure, closes on most inhibitors upon binding. In other cases, the protein, upon binding an inhibitor, has been shown to adopt an open position that is different from conformation of the apo structure.[61,62]

Ex20 is a BACE inhibitor from Janssen that contains a stereochemical center and two cyclohexane moieties each capable of adopting different conformations (see Figure 9). A data set of the complex was collected to 2.5 Å resolution and the structure solved by molecular replacement. After an initial

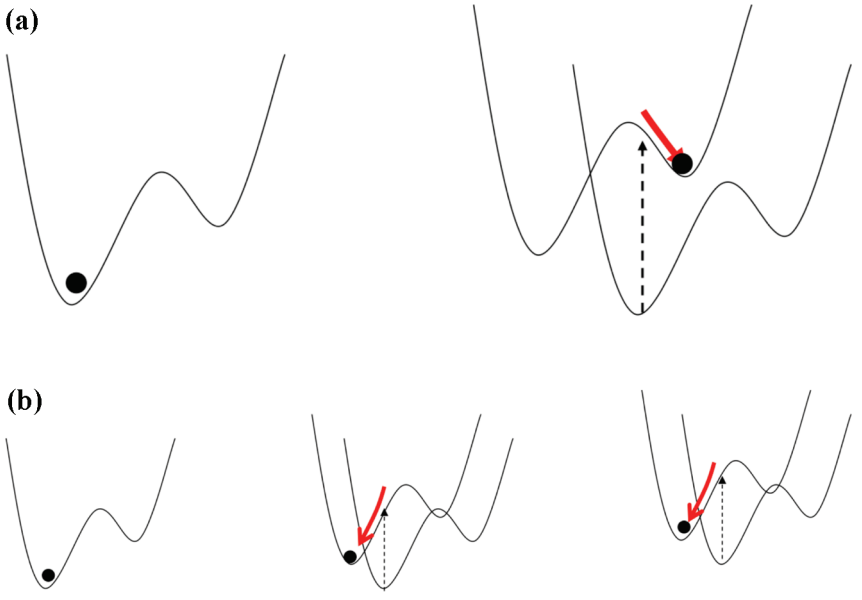

Figure 8. (a) Sudden addition of shape puts the molecule into a different energy well, i.e., conformer, leading to a nonadiabatic transition. (b) Gradual addition of shape ensures that the molecule remains in the same conformation and gradually is modified to fit the electron density.

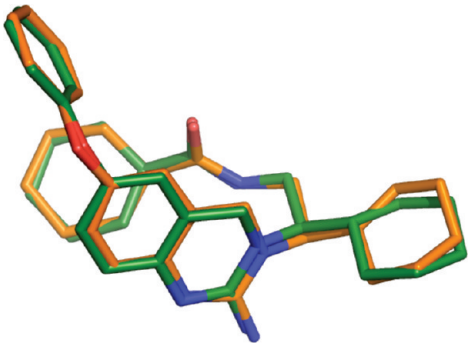

Figure 9. Superposition of initial and final positions of the ligand (green, AFITT; orange, refinement).

round of refinement, additional density was clearly observed in the active site.

In the first AFITT stage the blob found in the active site was smaller than the ligand. Visual inspection of the structure showed that side chain of flap residue Tyr71 ought to be rotated from the apo structure starting position to a new position away from density present in the binding site (see Figure 10a and Figure 10b). Once the position of the tyrosine was modified, the program readily identified a blob of density the size of the ligand (see Figure 10c).

As a test, we chose to enumerate the stereochemistry. The program suggested one conformation for each stereoisomer, with a clear preference for the *R*-stereoisomer (the details are found in the Supporting Information Table 1); i.e., the crystallographic data were used to successfully regenerate the correct stereochemistry. AFITT placed both cyclohexyl moieties in minimum energy conformations compatible with the electron density.

Macromolecular X-ray refinement algorithms require the generation and use of a force field parameter file for any small molecule ligands, and AFITT generates such a file based on MMFF94[63−66] parameters. The protein−ligand complex structure (PDB code 2wjo) was readily refined in Buster[67−70] to an R_{free} of 24.0% and an R_{factor} of 19.2% using this small molecule parameter file (see Table 2 in the Supporting Information for details).

In this complex, the flap adopts an open conformation similar to that observed in the apo protein. The only difference in conformation occurred for the side chain of Tyr71. The rotation of Tyr71 created a hydrophobic pocket comprising residues Val69, Tyr71, Trp76, and Phe108, into which the phenyl moiety of Ex20 fits snugly. The observed binding is in excellent agreement with previously reported structures for other compounds in the series.[71]

Retrospective Case. In a seminal paper by Perola and Charifson[72] the authors showed that about 10% of the ligand conformations in 150 complex structures, 100 public and 50 private, had unacceptably high conformational strain. Using two different sets of starting coordinates for the ligands, we rerefined 39 of the public structures for which electron density data sets were available. The first set used the deposited coordinates for the ligand during rerefinement. The second set used coordinates generated by AFITT; i.e., each ligand is removed, an unbiased electron density is formed from the structure factors, and the procedure described above is applied. Each ligand starts as a SMILES string, and so has no memory of its original crystallographic coordinates. For each set we then calculated the strain energy of the local minima,

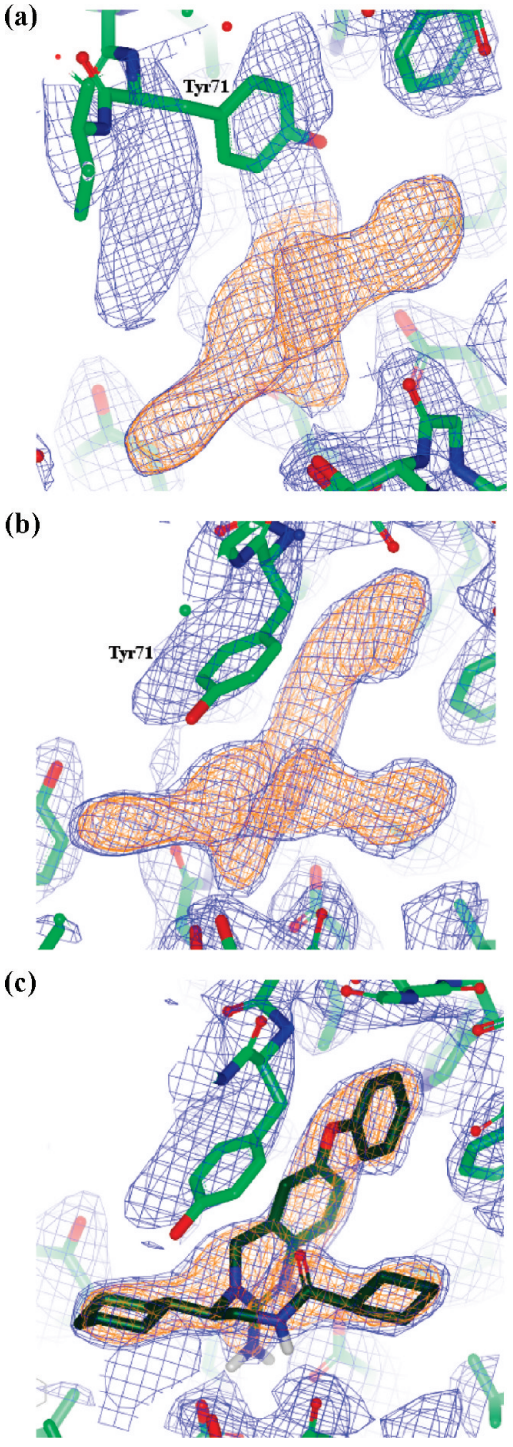

Figure 10. (a−c) View of the active site. Initial density after the initial Buster round (done with apoprotein) is in dark-blue. Selected "blob" is in orange. (a) Initial blob, prior to rotation of Tyr71. (b) Blob found after rotation of Tyr71. (c) Same as in (b) plus initial position of the ligand.

using MMFF and the electron density as a constraint during minimization to keep the ligands in the same conformational state. (Perola and Charifson used a similar method but with coordinate restraints). In all cases the deposited coordinates and AFITT generated coordinates had an equivalent fit to the data as measured by the R_{factor}. However, Figure 11 shows that, with one glaring exception, AFITT generates a conformation that is equivalent or lower in geometric strain when

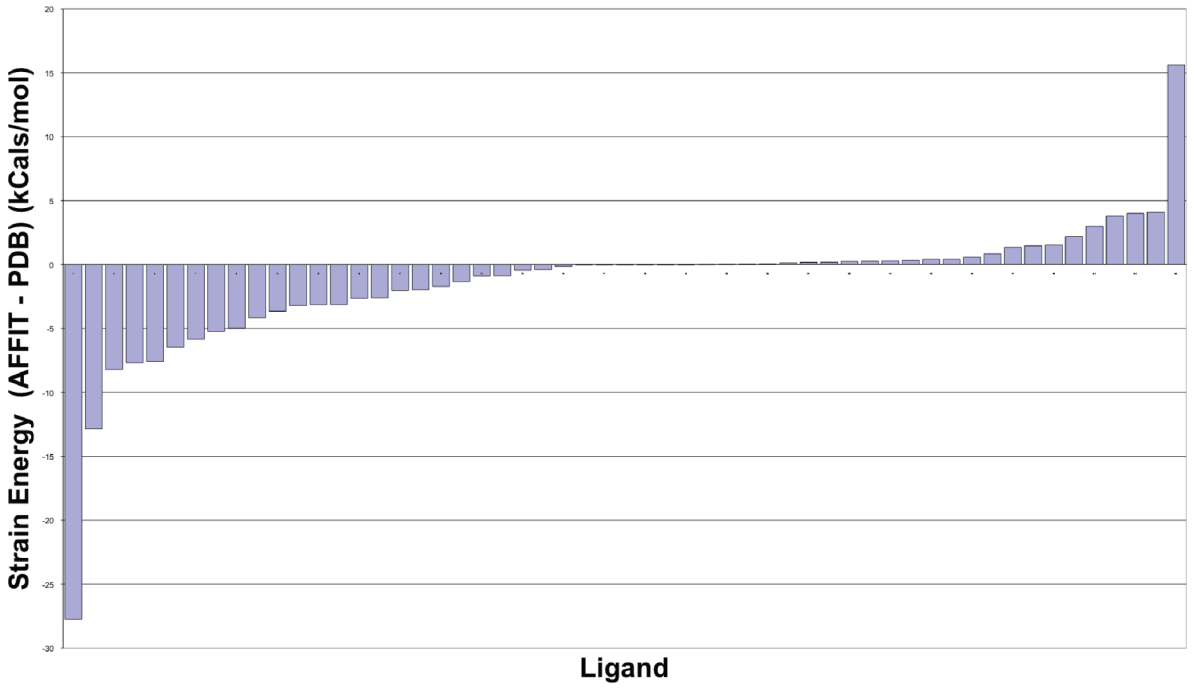

Figure 11. Difference in local conformational strain between rerefined ligands starting with an AFITT generated conformation versus the deposited conformation.

compared with the rerefined deposited coordinates. In this one exception there are two copies of the ligand, one of which AFITT handles well and the other for which it fails to find a low-energy conformation. This work was sufficiently encouraging that a project is currently underway over a much larger number of structures.

In conclusion, we believe a shape-based approach to crystallography has significant advantages. It is both intuitive and easy to apply and seems to result in ligands of lower intrinsic strain than traditional methods. An interesting future direction is the application to fragment-based drug design where crystallography plays an increasingly important role. For instance, when pools of fragments are applied to a crystal, careful consideration of the distinctiveness of the shapes within each pool may enhance the ability to resolve binding events. With the shape concept in mind, many other applications in crystallography are likely to arise.

Pose Prediction

A number of tasks in structure-based drug design rely on the availability of high quality models of the small molecule of interest in the context of its macromolecular target. When available, experimentally derived structures are highly prized, as they represent models that contain the highest quality information about the detailed interactions between protein and ligand. However, during the course of a structure-based design campaign, it is not reasonable to expect that every compound in a series will have its three-dimensional structure determined. This presents a great opportunity and a challenge for molecular modeling: to accurately generate useful models of the pose of a ligand in the protein active site using available experimentally derived information.

Quite a few approaches have been developed for bound-pose generation. At its most challenging level is the "docking problem". The docking problem is a difficult and currently unsolved problem in the field. To succeed, docking requires the ability to recognize the correct binding pose from a potentially large number of alternative but probable poses in the active site pocket. This is notoriously challenging, and there is currently no scoring function that can do this robustly. Generating the bound pose, without the ability to recognize it, is of little use.

Other methods attempt to leverage structural data by incorporating additional experimental data. For instance, by incorporating the positions of known ligands located in the binding pocket, one can potentially gain superior information pertaining to prospective research. For example, established methods utilizing ligand-only information to build a "pharmacophore" can benefit significantly from using structural information of both ligand and protein.[73] In its most general sense, a pharmacophore is an attempt to extract steric and electrostatic features that are common to a ensemble of atoms, whether ligand or protein, that are directly involved in bound interactions.[74] To be of use to medicinal chemists, a pharmacophore model needs to navigate a fine line between being generally applicable and just encoding the specificity of known ligands. Extrapolating from the known to the unknown also requires careful alignments in order to organize the data in a useful manner.

Alternative classes of pose generation strategies are the hybrid 2D/3D graph-based methods for aligning query molecules onto experimentally determined small-molecule structures. This is roughly analogous to homology modeling in proteins, with the template in this case being a bound conformation of a known small-molecule inhibitor. A number of template-based approaches to pose generation have been reported in the literature, and we mention only a few here. The CORES (complexes restricted by experimental structures) method[75] breaks up bound ligands into core fragments, which are then used to guide the docking of conformers of new compounds that share common fragments. Methods based on extensions to the concept of maximum common substructure[76] have appeared recently in the literature, for example,

the maximum overlapping set[77] that attempts to account for common (but nonconnected) fragments between molecules. A similar approach, termed graph-based molecular alignment,[78] uses a more sophisticated optimization procedure to more robustly bring molecules into common alignment.

The template methods described above all rely on an atomistic representation in order to drive pose-generating overlays. This has innate limitations; e.g., exploration of new molecules must have closely matching atoms in very similar arrangements in order for the overlays to succeed. A more flexible approach employs molecular shape similarity to more generally leverage knowledge of bound conformations. At a basic level, the method attempts to mimic not atom positions and identities but rather the fundamental shape of the bound pose. In this way it moves away from the rigidity of prescribing atomistic arrangements.

There are many possible realizations of this approach. One straightforward method is to pregenerate a set of conformers, then overlay this set of conformations using shape, and then rank order the overlays based on shape similarity.[79] We have taken this approach on a set of 340 protein−ligand structures across 28 different targets from the Lilly data set originally constructed for studying docking and posing.[80] For each structure we generate a SMILES string for the bound ligand, then input this into Omega[81] (version 2.3.2) to generate a set of conformations, followed by shape-driven overlay onto the cognate ligand conformation as in ROCS. The performance of this method is shown in Figure 12a (red squares) and can be seen to be effective and is an easily implemented procedure. We refer to this method as "Omega-ROCS".

An improvement on the Omega-ROCS approach is the topic of the present study and represents a more robust method for generating shape-driven poses we have called Shapefit. The inspiration for this method is the fitting approach to crystallographic refinement used by AFITT, as described above. In the case of Shapefit, the density being fit is the molecular Gaussian function produced by the bound ligand, which could be thought of as a type of ideal electron density. Comparison of the performance between Omega-ROCS (red squares) and Shapefit (blue triangles) is shown in Figure 12a. As expected, the results from Shapefit are systematically superior to using Omega-ROCS in fitting the ligand to itself, albeit from conformations generated from its SMILES representation.

However, reproduction of a known cognate ligand conformation is not the primary motivation for this work. We wish to use experimental knowledge of bound ligands, in the context of the host protein, to leverage *prospective* predictions of bound poses with some measure of a degree of confidence in the output. Thus, we require a way to quantify our proximity to what is known experimentally and use this as a measure to report quality in the prediction. To do this, we selected a set of 363 ligand pairs across the 28 representative targets and generated a bound pose using the Shapefit calculation on progressively dissimilar pairs of ligands for each target. In this way we attempted to gauge the decline in effectiveness as the similarity between the two molecules decreases. Shown in Figure 12b are the results of this analysis. It can be seen that as the shape and color similarity, as measured by the Tanimoto Combo, between the docked molecule and the bound (shape guide) molecule decrease, the range of rmsd values, as calculated between the docked molecule and its known bound conformation, increases. Tanimoto Combo here is simply the sum of the Shape Tanimoto, defined above, and the Color Tanimoto, where the latter is similarly defined as the overlap of "color" Gaussians, i.e., representing chemical functionalities, divided by the difference between the self (color) overlaps and the color overlap of the two molecules. A Tanimoto Combo cutoff of around 1.4 is a good threshold above which reliable results are obtained.

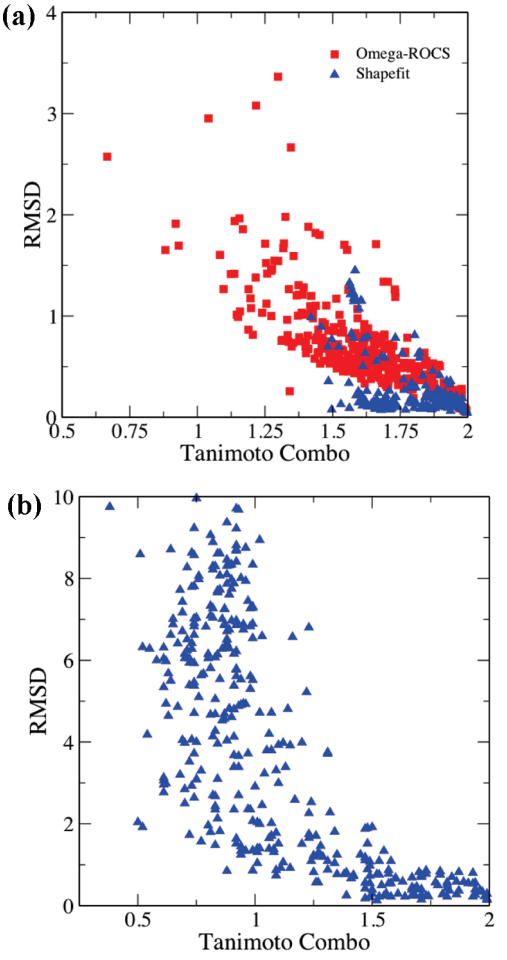

Figure 12. (a, b) Here, 340 crystallographically determined protein−ligand structures were used, from 28 different protein targets. The first graph shows the fitting of each ligand to itself, using either ROCS (red squares) or the Shapefit (blue triangles) method. The Omega-ROCS method uses Omega[110] (version 2.3.2) with max conf = 30 000, energy window = 100 kcal/mol from the SMILES string of each ligand followed by overlay with ROCS[111] (version 2.3.1); the Shapefit method takes each such alignment and further optimizes the Tanimoto Combo score. The Tanimoto Combo and rmsd are calculated for the highest scoring pose against the crystal structure. Part b shows Shapefit overlays for 363 ligand pairs, chosen randomly from within target series. In this plot the rmsd is calculated as before but the Tanimoto Combo is calculated between the two (different) overlaid molecules. A cutoff of around 1.4 in the Tanimoto Combo is sufficient to provide a high degree of confidence in the heterologous pose prediction.

We examined two specific examples of ligands with bound conformations that proved challenging for the Omega-ROCS method. Shown in Figure 13 is an extended "bent-back" conformation of a bound inhibitor whose conformation has been determined by X-ray crystallography. Figure 13a shows the overlay produced by Omega-ROCS (cyan-colored carbons), and Figure 13b (pink-colored carbons) shows that for Shapefit. The X-ray crystallographic pose is shown with green carbons in both figures. It can be seen that the overlay produced by Shapefit is of significantly higher quality.

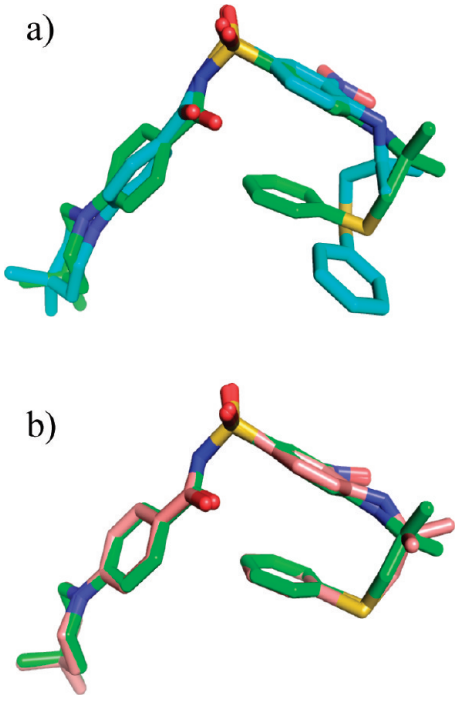

Figure 13. (a, b) Example showing performance of two overlays for the Omega-ROCS and Shapefit methods. The X-ray crystallographically determined conformation is shown with green carbon atoms. (a) Omega-ROCS generated pose, shown with cyan carbons. (b) Shapefit generated pose, with pink carbons. The robustness of the adiabatic fitting algorithm in Shapefit produces systematically superior overlays.

Finally, in Figure 14 we show another example of a comparison between the overlays produced by the two methods. As with Figure 13 the X-ray crystallographic ligands are shown with green carbons. In Figure 14a is shown the overlay produced by Omega-ROCS, and in Figure 14b we show the Shapefit produced overlay. For both examples Shapefit is the superior method.

Our conclusions are that using Omega to pregenerate conformations followed by overlays performed by ROCS gets you close but is not sufficient to drive the ligand overlay to a pose sufficiently equivalent to the cognate pose for practical, prospective use. We have illustrated this by comparison to a more robust method, which uses an adiabatic algorithm to couple force field to shape. The Shapefit methodology seems to be systematically superior for performing pose generation when one possesses knowledge of existing bound conformations of similar molecules. There are still outstanding issues here; for example, how do we use multiple ligands in this process? Do we just use the one closest in shape and color space or a weighted combination of all the results? To what extent can we use binding information from homologous proteins? Can we use a combination of shape matching and graph-based methods to better effect? What we believe we have shown is a clear demonstration of the value of shape information in pose prediction; future work will inevitably increase our understanding of how to use it most effectively in structure-based design.

Library Design

Modern library design increasingly adopts rational strategies intended to maximize the biological relevance of molecules. This results in relatively small libraries and an emphasis

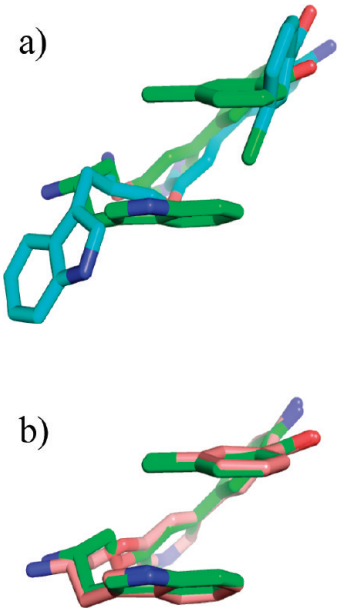

Figure 14. (a, b) Another example comparing the performance between Omega-ROCS and Shapefit methods. The X-ray crystallographically determined conformation is shown with green carbon atoms. (a) Omega-ROCS generated pose, shown with cyan carbons. (b) Shapefit generated pose, with pink carbons. The robustness of the adiabatic fitting algorithm in Shapefit produces systematically superior overlays.

on populating screening collections with many different scaffolds.[82] Given the importance of the shape of a molecule in molecular recognition, a rational strategy would be to incorporate control of shape variability into the design of libraries, for instance, as has been commented, to avoid collections that are too "flat".[83] The approach at AstraZeneca has been to define the "shape-space" we want to cover by using reference shapes obtained by using a simple clustering procedure previously described.[84] In short, the shape similarity to an initial reference structure is calculated for a set of molecules and the most dissimilar is chosen as the second reference shape. Shape similarities are calculated to this second structure, leading to a third reference structure that is least similar to the first two. The procedure continues until the level of minimum similarity remaining to all reference structures is higher than a cutoff criterion. The number of reference shapes so obtained is very dependent on this value. For our purposes, a suitable level of discrimination is an ST[a] of 0.75.[31] This level of similarity is sufficient both visually (important for chemist buy-in) and empirically for virtual screening. Figure 15 illustrates the strong (exponential) dependence of the number of reference shapes on heavy (non-hydrogen) atom count (HAC). This makes the extension to larger molecules problematic. However, a common feature of molecules designed for screening is that they should have a reduced complexity.[85] One of the simplest ways to ensure low complexity is to limit the molecular weight range[86] or alternatively the HAC range. A useful range for leadlike libraries is a molecular weight less than 450[87] that is approximated by molecules with 26 or fewer heavy atoms. Figure 16 shows examples from the 5408 reference shapes produced from an analysis of one million

[a] Abbreviations: ST, Shape Tanimoto; PPI, protein−protein interactions; HAC, heavy atom count; USR, ultrafast shape recognition; SM, steric multipole; OMI, overlap from moments of inertia.

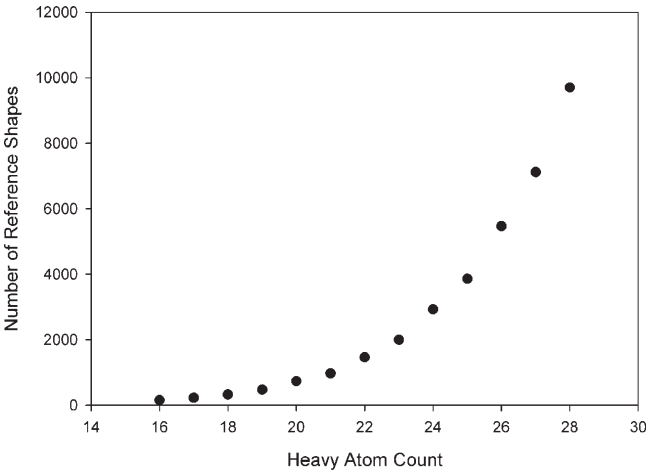

Figure 15. Number of reference shapes generated when clustering molecules grouped by heavy atom count. The plot is extended from 18 to 26 heavy atoms to show that the behavior remains smooth.

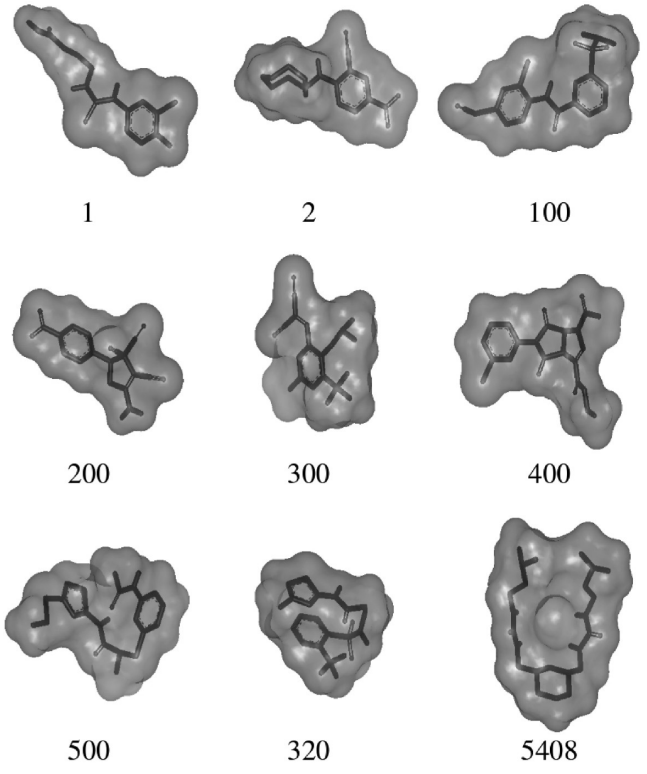

Figure 16. Reference shape examples from a 1 million molecule library (HAC between 12 and 26).

commercially available compounds with a heavy count (HAC) range of 18−26.

The shape clustering method is designed to ensure that each of the one million molecules is "close" to at least one of the reference shapes. The accuracy with which shape-space is covered is tested by assigning a different set of two million commercially available compounds with the same HAC range to the reference shapes. We found that only 0.1% of these molecules are more dissimilar than 0.75 ST to any reference shape. This confirms that the reference shapes are a good representation of the shapes of molecules in this HAC range. Molecules may actually be closer than the cutoff criterion to more than one reference shape; in fact on average each molecule is close to about six reference structures. However,

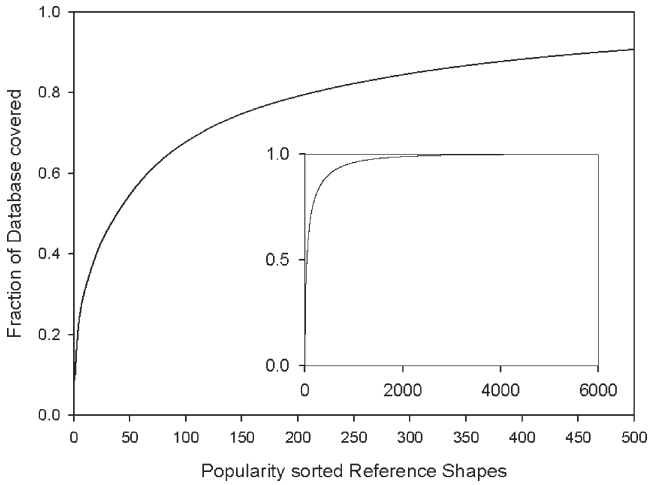

Figure 17. Coverage of the 3 million molecule database by the top 500 reference shapes after they have been sorted by popularity. Inset is the coverage extended to include all 5408 reference shapes.

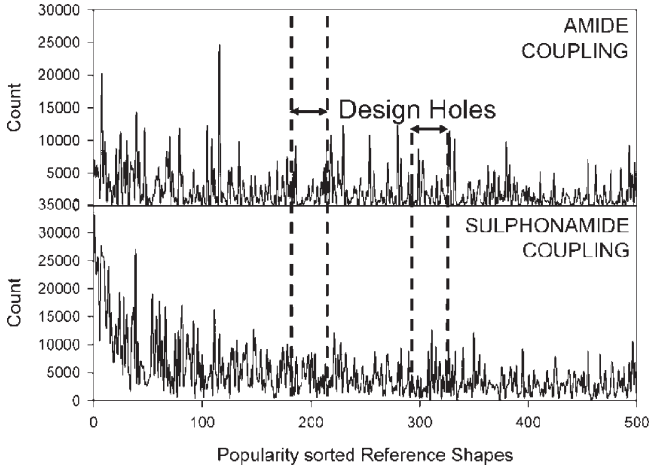

Figure 18. Shape profiles of a library formed from a simple amide coupling reaction (bottom) and a sulfonamide coupling reaction (top). Two design holes in the amine-coupling library are well filled by the sulfonamide coupling.

some reference shapes "cover" more molecules than others and this leads to the concept of shape "popularity". This popularity is quite nonuniform, and a relatively small subset of shapes are close to a large proportion of molecules, reminiscent of the dominance of few common frameworks in analysis of chemical scaffolds.[88,89] In fact, as shown in Figure 17, 500 of the most popular reference shapes cover about 90% of the 3 million molecules. As such, these 500 reference shapes define a practical method for quantifying shape variability in library design.

One strategy for designing a screening collection is to ensure an approximately uniform distribution of molecules among the reference shapes. This is difficult to achieve without some aspect of design because of the "power-law" distribution of the popularity of the reference shapes; i.e., a simple random selection would be biased toward just a few of the popular shapes. Figure 18 shows a profile of a library formed from a simple coupling reaction to produce amides. The profile has the expected characteristic of a few well-populated shapes. However, it also shows that even this simple reaction produces molecules with a good degree of shape variability, which by careful selection could be used to augment a screening

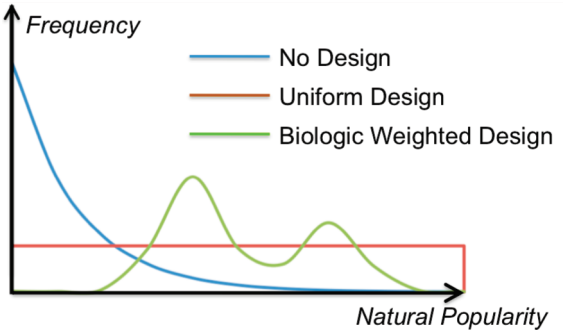

Figure 19. Possible design strategies for shape libraries. The natural abundances of shapes can be adjusted to be uniform or to reflect biological knowledge as to the best shapes screen.

collection, to improve its overall shape profile. Careful inspection also shows that there are specific gaps in the shapes of this library, for instance, reference shape 320. In contrast, a profile of a library in which the same reagents are coupled to produce sulfonamides, populates shape 320 as shown in the figure. The two chemistries complement each other in the construction of a diverse shape library.

Two other design strategies can also come into play. The first is to include other notions of diversity, such as those based on chemical similarity, or other molecular properties, i.e., such that molecules similar to a given reference shape populate an "orthogonal" axis. One appealing property that complements the simplicity of shape is the electrostatic potential. Initial experience with such suggests the diversity of potential maps is as expansive as shape itself, making complete coverage difficult. The second is to deviate from uniform coverage toward biological activity; i.e., it may be that some shapes are preferred for activity. This is illustrated in Figure 19 but is currently still speculative, in particular because the past history of compound construction may bias our findings. Current research is aimed toward exploring these questions and the underlying shape space of druglike molecules.

Binding Site Shape

KOMPAS, a joint project between GSK and OpenEye, was designed to generate negative images of active sites for use in virtual screening when the native ligand is unobtainable or unrepresentative of the desired small molecule. The mathematics and implementation of this rely heavily on the formalism developed by Pickup and Grant.[58] In that work the authors show that molecular volume, as defined by a set of fused spheres, can be accurately represented by a set of atomic-based Gaussians, an observation that gave rise to ROCS. They further showed that the derivative of the function formed from this set of Gaussians, with respect to atomic radii, has the character of an area. Considering the approach of two atoms, they reasoned that although the overlap of representational Gaussians increases, this area term decreases. As such, with appropriate parametrization, a linear combination of overlap and "area" can simulate the repulsive and attractive terms from van der Waals forces but with the familiar advantages of Gaussian functions. This Gaussian-equivalent of dispersion forces has been used in docking;[90] here it is used in a shape context to define small volumes of high contact potential between a probe carbon atom and the protein on a regular cubic lattice. Contouring of this grid produces "cores", small, distinct volumes that form the basis of putative molecular shapes. Selection of a set of proximal cores followed by convolution by a sphere or Gaussian representation of a carbon atom expands these cores into a ligand-like shape. An example of this process is shown in Figure 20. These shapes can then used as the starting points for shape-based virtual screening.

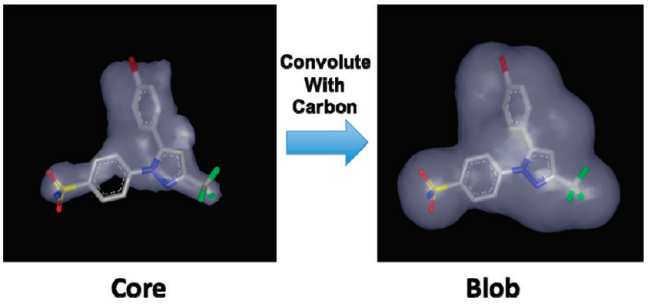

Figure 20. Illustration of shape generation in KOMPAS. On the left is a contour of a Gaussian contact function for 6COX. This is not derived using the ligand. On the right is the result of convoluting this shape with a Gaussian the volume of a carbon atom and recontouring.

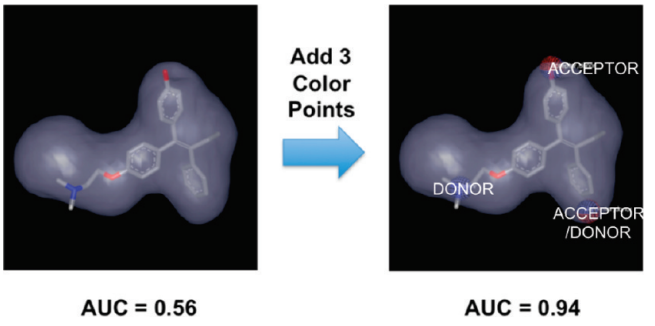

Figure 21. Illustration of the effect of adding "color" points to the active site shape generated for 2ERT. Just three points describing likely interactions with the protein, added by graphical editor, rescue the poor performance of the raw shape.

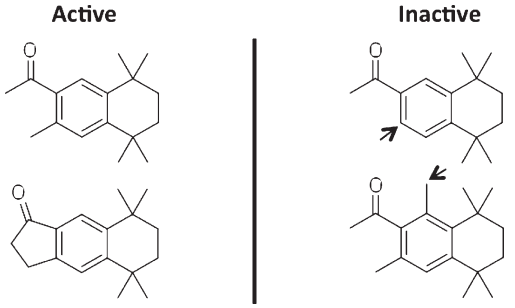

Figure 22. At left are two small molecules with specific musk odor. At right are two molecules that differ by only a single methyl group from the top musk but that exhibit no musk odor. Musk activity is assessed by professional "noses" and quantified generally as strong, weak, or absent. Activity is driven by specific agonism of GPCRs in the olfactory neuroepithelium; the ligands above were part of the study that introduced the Compass[96] technique. Inactivity of the top nonmusk may be due to inability to activate an allosteric switch in the olfactory GPCR responsible for musk perception; it cannot be easily explained by a decrease in binding affinity. The bottom nonmusk may simply be too large for the active site, but there may be other mechanisms. In making this distinction, it is clear, however, that the issue comes down to shape in quite a pure sense, since the molecules are rigid.

Prospective Example. Protein−protein interactions (PPI) are challenging targets for small-molecule drug discovery.[91,92]

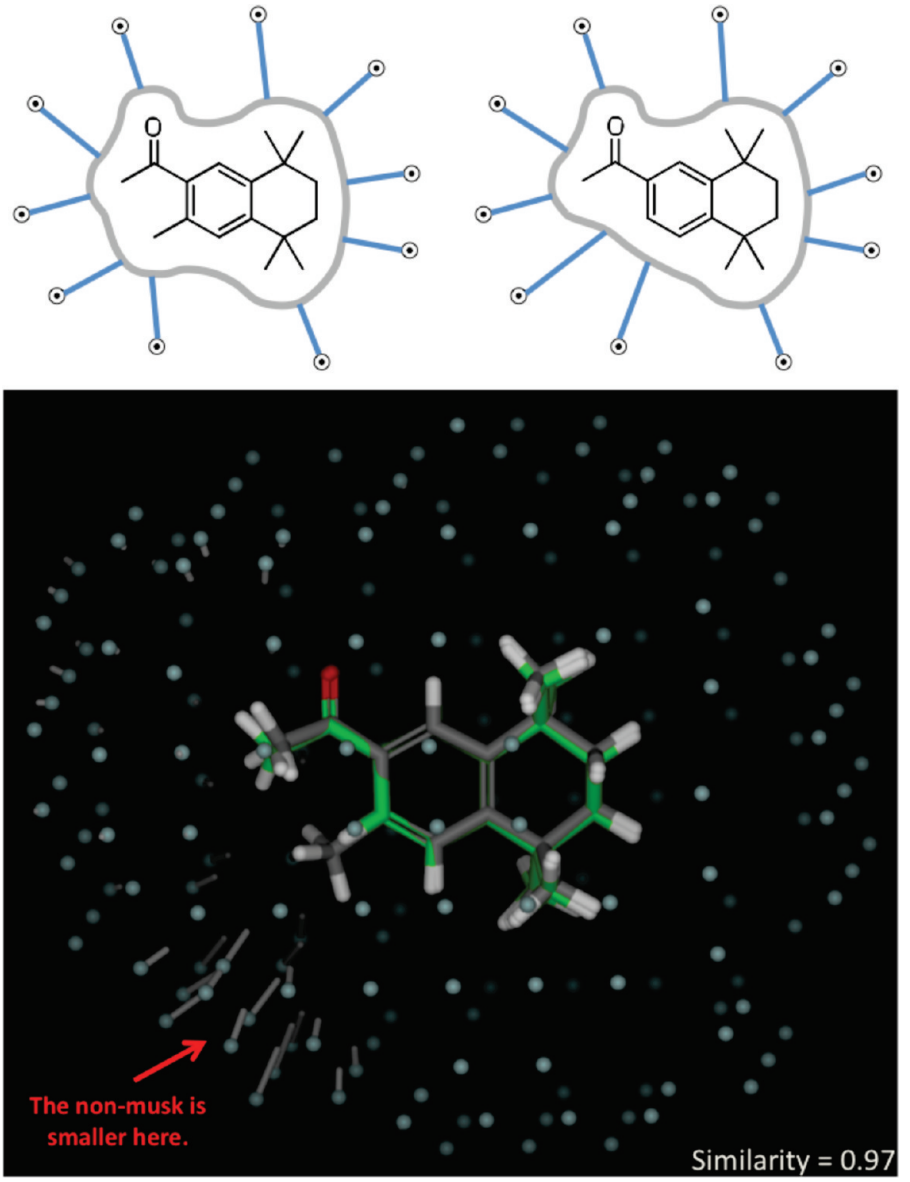

Figure 23. Molecular shapes can be characterized by the distances to the molecular surface from points in space. The differences in these distances form the basis for comparison between molecules. At top, a musk and nonmusk are cartooned with distances from observers placed outside their surfaces. The corresponding distances are longer on the lower left, where the methyl group is missing. This is depicted in 3D, with the differences in distances shown as rods emanating from observers. By use of a normalized Gaussian function of the distance differences, a similarity function can be defined whose optimum rewards surface concordance but yields a limited penalty for discordance. The function is continuous and piecewise differentiable with respect to molecular pose, which has advantages for optimization of relative poses.

Even when PPI crystal structures are available, finding good small molecule leads for chemistry optimization can be difficult. Rational design starting with the protein or peptide partner is one option, but this approach is often not ideal. The binding site can be shallow and/or dominated by a few key interactions of distal residues. If the designed molecule is too peptidic, pharmacokinetic issues will arise. Consequently, many rounds of optimization are required to develop a more druglike molecule. Thus, synthesizing peptido-mimetics may not be the most straightforward route. Furthermore, the natural protein/peptide partner may not occupy the full binding site, and because of the shallow nature of the pocket, a small molecule inhibitor maximizing every available interaction in the site is most ideal.

KOMPAS was used toward this end in a current PPI project at GlaxoSmithKline (GSK). Filtering by shape volume and site depth resulted in three shapes for three different sites, which appeared by inspection to be drug-sized and deep enough to potentially inhibit the PPI. The peptide bound to the protein target passed through each of these three sites. ROCS virtual screens against the GSK compound collection were then carried out for the three shapes, and the top 10 000 shape-matched compounds per site were retained and filtered with the Gold, version 3.2, docking program using the high-throughput settings and ChemScore[93,94] scoring function for each site. Top scoring poses were then visually inspected within their respective binding site. From the shape virtual screening exercise, 200 compounds were screened in a cell reporter assay resulting in five hits representing three chemotypes for two of the three identified sites. Later, a high throughput screen (HTS) was run utilizing a binding assay designed to measure disruption of the protein−peptide

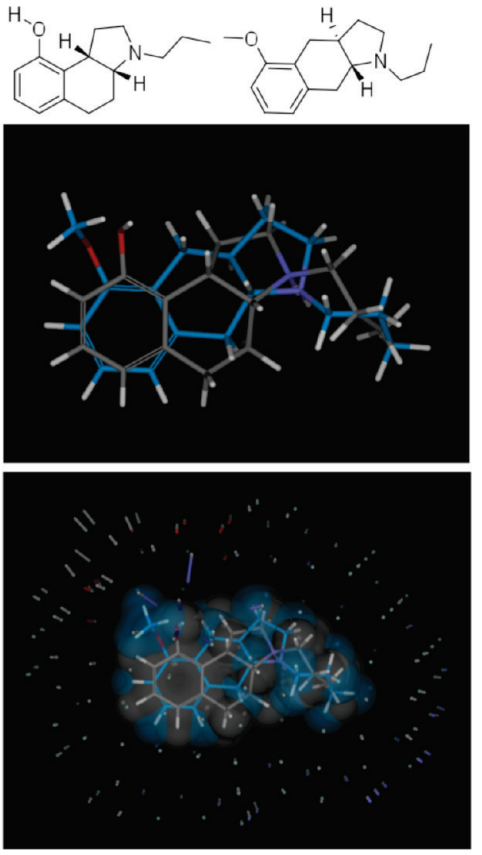

Figure 24. Two serotonin ligands with differing underlying scaffolds are shown. The concept of observers and distances generalizes to polar surfaces by measuring minimal distances (and directions) to charged atoms. The highest scoring mutual alignment is shown at top. Despite significant differences in the underlying scaffolds, the procedure is able to identify joint poses where the steric envelopes are remarkably similar, the charged amines are tightly aligned, and the oxygens of both ligands are able to accept hydrogen bonds from the same part of space. The differences are very minor (bottom panel), primarily resulting from the change from hydroxyl to methoxy, resulting in the gray difference rods (steric differences) and the blue difference rods (indicating a missing donor on the ligand with blue carbons). Other differences are minor, with very slight differences in position and orientation of the acceptor functionality of both molecules. Overall similarity is 0.82 (scale of 0−1).

interaction, and compounds with the aforementioned chemotypes were active with pIC_{50} values in the 4−5 range. The pIC_{50} for the best HTS hit was 6.5, indicative of the challenge of finding potent inhibitors of PPIs via HTS. Orthogonal biophysical assays ruled out two of the three chemotypes, but one shows weak binding by surface plasmon resonance (SPR) studies. A small set of compounds was synthesized by the GSK Exploratory Chemistry group to generate SAR based on the predicted binding mode generated from the shape/docking protocol. For the chemotype with activity in the binding assay and SPR, removing the key predicted pharmacophore elements eliminated all activity and SPR binding.

Retrospective Example. Defining active site volumes is an imprecise science, and using just shapes as queries for virtual screening provides a relatively porous filter. In PPI this is not so important because one is typically looking for the unusual, and so the lack of restriction can be an advantage. However, there are often clear interactions with the protein that could be captured in addition to shape that would aid in discovery. A retrospective example illustrating this from the DUD data set[95] is shown in Figure 21. The three interaction, or "color", points are added using a graphical editor. This consists of selecting several points on the surface of the contour and averaging their position for each color point to be added. They are then further defined by their hydrogen-bonding character and were selected on the basis of knowledge of the protein and known ligand, although it should also be possible to just use the protein in a prospective manner. A shape, or shape plus color points, is then used as standard input to the ROCS program; i.e., conformational expansions of known ligands and decoys, in this case generated by Omega (version 2.3.2), 100 conformations at most per molecule, are scored based on their best overlap to the input and standard statistics for recall calculated. As can be seen, just a few judiciously chosen color points can make a startling difference in recall rate, as measured by the area under the curve (AUC) for the DUD actives and designed decoys for this target. Shape alone does not distinguish at all between the decoys and actives, while the addition of chemical typing gives close to a perfect separation between classes.

Reinvestigation of the PPI target using such added information is currently underway. The potential to annotate shapes also may overcome one of the disadvantages we found in generating such shapes when intense electrostatic interactions are prevalent. KOMPAS shapes are constrained to average contact potentials, and areas of strong electrostatic complementarity often have van der Waals interactions beyond this normal range; i.e., they do not appear in the core shapes. Changing the Gaussian parameters to extend the core domains can ameliorate this, but this then causes clashes with the more hydrophobic parts of the protein. A more general solution that either incorporates both types of contact shape or that automatically annotates the hydrophobic shape with donor and acceptor sites would go a long way to solve these problems.

In summary, it is unclear if the chemotype identified by the shape generation/docking protocol will be incorporated into an inhibitor molecule for this PPI program and studies are still ongoing. Nonetheless, shape technologies appear to be promising tools for representing a virtual molecule in a binding site, especially for PPI targets where every $Å^2$ of binding site surface area can be critical for affinity. In this example, the program team believed in the shape concept and experimentally screened molecules selected by this shape-based virtual screen. At a minimum, we now better understand the functionality we desire from shape-based computational tools and look forward to progress in this area. Finally, application of this technology to protein−protein comparison is straightforward, while the potential to apply the shape concept to polypharmacology and to specificity prediction remains to be tapped.

Shape from Surfaces

The intention here is to discuss a slightly different approach to that described earlier in this article and that has been developed with the goal of describing molecular shapes accurately as *surfaces* with properties such as polarity.

Whereas many representations can be thought of as 3D, subtle surface differences are challenging to capture, but subtle differences can have large effects. Figure 22 shows the striking effects of addition or deletion of a single methyl group to render a simple and relatively rigid molecule devoid of activity mediated by an olfactory GPCR. Figure 23 shows an

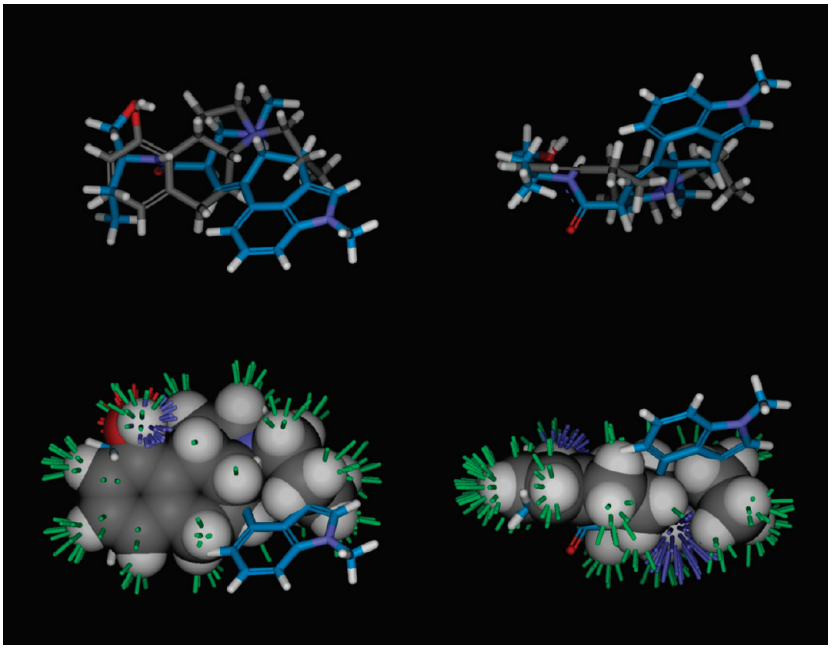

Figure 25. Surface comparisons of this type can be highly effective for virtual screening. Here, a competitive serotonin ligand (methysergide, shown with blue carbons) was identified in a virtual screen at a false positive rate of less than 3% against a set of druglike screening compounds. The size difference between the ligands would pose a challenge for approaches that rely upon concordant volumes or upon pose optimization strategies that penalize mismatches instead of rewarding matching features. The similarity of methysergide to the angular tricyclic serotonin ligand was 0.73, reflected in the degree to which the similarity sticks populate the surface of the target ligand. The right-hand panels attribute similarity to the atoms of the angular tricyclic compound (gray carbons), with green rods indicating high shape similarity, red indicating high similarity for negative polar moieties, and blue indicating high similarity for positive polar moieties.

approach to shape characterization that is capable of discriminating fine changes in molecular structure in a scaffold-independent way. The approach characterizes molecules as collections of distances from observation points in space. The collection of minimal distances to a molecule from a fixed set of observers encodes a packing of spheres that graze the surface of a molecule, forming essentially a perfect binding pocket around the molecule in question. Differences in the radii of the spheres can form the basis for comparison of molecular shapes. This basic representation underpinned the Compass 3D QSAR approach[96−98] (no relation to the Kompas project described above).

In the development of this approach, construction of virtual binding sites entailed two complications. The first was how to address the question of *which* ligand poses were relevant to biological activity. This was addressed by developing a formalism for machine learning wherein one could simultaneously choose poses of ligands while estimating the parameters of the model for the ligands' activity. An elegant formalization of this early work on musks, called multiple-instance learning,[99] has found applications in many areas of machine learning. The second complication was how to address the issue of molecular surface polarity. A straightforward extension of the distance-measuring concept from Figure 23 is to measure different sets of distances, each to different sets of atoms. One set of distances corresponds to the minimal distance to *any* atom. Another set corresponds to the minimal distance to any hydrogen bond acceptor or formally negatively charged atom. The last corresponds to distances to donors or positively charged atoms.

With the addition of directionality (the degree to which an observer and a particular atom are compatibly oriented) and charge magnitude, a generalized function of molecular similarity was developed[100] and subsequently generalized to specifically address virtual screening considerations.[101] Figure 24 shows the application of the approach to identifying an optimal joint superimposition of two competitive 5HT1a ligands. Underlying scaffolding is unimportant, with both ligands exhibiting remarkably similar surfaces, in terms of pure shape and the disposition of polar moieties. Such joint superimpositions may be used as the target for virtual screening, much as one uses protein binding pockets for docking. Figure 25 shows the result for methysergide, a 5HT1a ligand of very different structure from those that formed the screening target. The advantage of constructing a similarity function that *rewards* surface concordance (as opposed to penalizing deviation or computing volume overlap) is that a ligand such as methysergide, which extends beyond the envelope of the known ligands, can be retrieved at a very high rank against a background of diverse, druglike screening molecules.

This conceptualization of molecular similarity is quite general, and it can be applied to comparing protein surfaces. Figure 26 shows an alignment of CDK2 and c-Met, the former being a cyclin-dependent serine−threonine protein kinase and the latter being a tyrosine kinase. The biological functions of these proteins is quite different, in terms of their substrates and their regulation, owing in large part to the very different overall exterior architecture of the two proteins. However, their ATP binding sites are actually quite similar, especially in the core region. Figure 27 shows a set of observers placed *around* the ligand of CDK2. The alignment of c-Met was optimized relative to the surface measurements made from these observers, resulting in the correct correspondence of the proteins. Despite their different biology, both proteins are effectively inhibited by staurosporine or close analogues. The pocket similarity can be quantitatively related to propensity to bind the similar ligands in the same fashion.

An interesting duality exists between proteins and ligands when one considers different methods for similarity

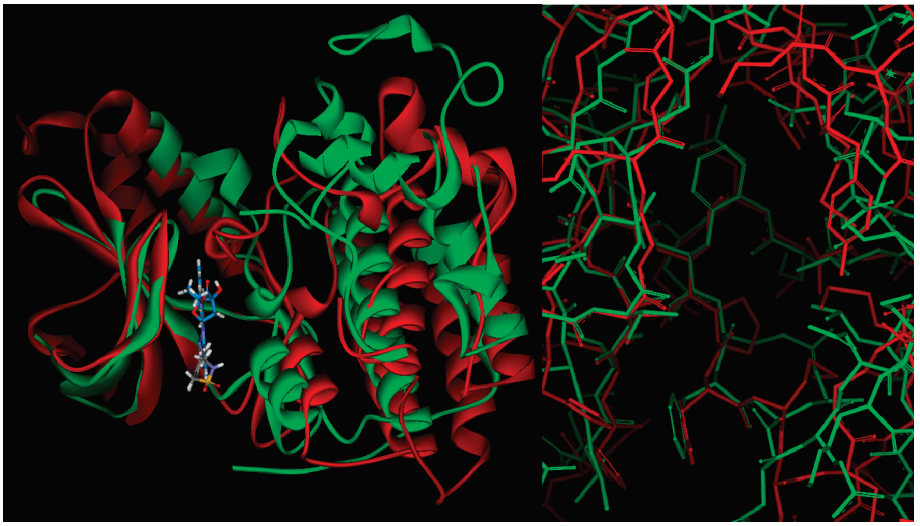

Figure 26. Two proteins are shown: CDK2 (1KE6, green) and c-Met (1R0P, red). They have modest sequence identity (less than 20%) and significant differences in overall structure at a global scale, especially in the right-hand lobe (evident at left). However, their binding sites (with bound ligands at left) are quite similar in structure (right). In the hinge region, c-Met makes use of a proline and tyrosine and CDK2 makes use of a glutamine and phenylalanine (blue arrows), but the surfaces are similar enough that both enzymes will bind staurosporine analogues in similar orientations, with analogous hinge binding interactions (hinge acceptor and donor are circled in yellow).

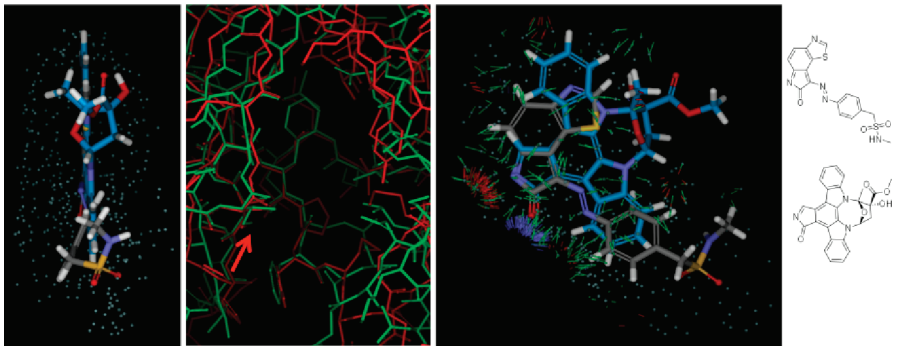

Figure 27. The protein alignment of c-Met to CDK2 was computed from the observers shown here outside the ligands (left panel). The right panel shows the relative alignment of the ligands (viewed from the right side of the left and middle panels). The analogous polar interactions of the two ligands (red arrow) manifest as an area of high similarity between the proteins. The overall binding pocket shapes are also relatively concordant (green sticks). The cognate ligand of the c-Met structure was closely related to staurosporine (blue carbons), which itself is a potent CDK2 inhibitor. The relatively high similarity in active sites between c-Met and CDK2 is exhibited both directly in the surfaces of their ATP binding sites and in the ligands that bind them.

computation. In the case of small molecules, we see a great deal of population of the space of ligands that are obviously similar to an existing known ligand in a 2D sense. We believe this is because 2D reasoning is a significant part of the design mechanism of man-made ligands.[102] However, when a target has been around for a long time, substantially more varied scaffolds are discovered, many of which are quite similar in a 3D sense to previously known ligands but which are quite different in a 2D sense. In the case of proteins, we see large families that represent small steps in terms of the mechanism of design (evolutionary steps of sequence modification). For ligands that have been in nature for evolutionarily lengthy time periods, such as ATP, we observe relatively conservative variations such as human CDK2, its species variants, and related proteins in very different organisms (e.g., CDC2 in yeast). We observe moderate jumps (e.g., c-Met as shown in Figures 26 and 27), where very significant local similarities relating to sequence are present. However, we also observe remarkably different binding sites that make use of ATP. Proteins such as phosphodiesterases (e.g., PDE4b and PDE5a) use a completely different architecture to make use of ATP, but the surfaces of the binding pockets can be correctly aligned using the approach to similarity discussed here.

Similarity metrics that are *mechanistically* related to a design process, 2D approaches for ligands and sequence-based approaches for proteins, are able to identify large numbers of functionally related molecules, but these are just the "obvious" set of relationships. By making use of shape, which is physically related to the *fitness* of a molecule for a purpose, we can identify both the obvious *and* the nonobvious relationships among molecules, both large and small.

Approximate Shape Methods

It is only the ever-increasing computational power of the past 20 years that has enabled meaningfully exact shape comparisons to be fast enough to be practical. But every increase in computational facility also increases the potential scope of problems to be tackled. So, for instance, although a ROCS-type calculation takes about a millisecond to compare two conformers, large enough libraries make this seem slow.

As such, faster but approximate methods might seem to have a role to play either as a prefilter to more exact calculations or as an adequate replacement. Other fields have faced similar issues; human fingerprints used to be matched laboriously by forensic experts. The development of feature reduction to "loops", "arches", and "whorls" allowed automated systems to screen millions, and given a sufficient number of feature matches, individuals can be identified with statistical confidence. Can the same be said of approximate shape methods? Here we have examined three candidates: our implementation of the ultrafast shape recognition (USR) method,[103,104] steric multipole (SM)[105], and finally overlap from moments of inertia (OMI). The OMI method is essentially ROCS without any of the time-consuming optimization of alignment, i.e., taking the overlap straight from two molecules aligned by moments of inertia. USR and SM generate a set of descriptors, 12 and 10, respectively, that are compared through sums of differences and as such are very fast, theoretically millions per second, whereas OMI is about an order of magnitude slower but still much faster than ROCS.

The graphs in Figure 28 and 29 illustrate the problems of these methods relative to the "exact" solution, i.e., a Shape Tanimoto derived from ROCS. In this test we have compared each target molecule in DUD to all the conformations of other actives and decoys, leading to about two million comparisons. Although each method has a correspondence with shape, the scatter illustrated in Figure 28 is considerable. In fact neither USR nor SM have much correspondence with shape. Why is this? One can learn much by examining the worst mistakes, as illustrated in Figure 29. Points in the upper left of each graph represent false negatives, i.e., shapes that are in fact very similar but that a method thinks are different. These are particularly prevalent with USR, and an example is given in Figure 29a. In general, false negatives are characteristic of "fragile" methods; i.e., small changes make large differences in the descriptor set. USR is particularly prone to these because of the sensitivity to the position of extremal atoms. If two atoms quite separate in space are each close to being the farthest from the molecule midpoint, a central part of the USR approach, then the removal of either may shift the definition of "extreme" and have a dramatic effect on the descriptor set. Figure 29b illustrates an example taken from the bottom right of the SM graph. Here the problem is one of false positives, i.e., molecules "masquerading" as being similar when, in fact, they are not. Steric multipoles are particularly sensitive to this because the multipole approach is weak on internal detail, capturing only the coarser aspects of shape. Hence quite different shapes may appear similar to the SM approach; i.e., the method is "blunt". Finally, the OMI method correlates well with the underlying similarity but illustrates that as the actual similarity gets worse, so does the variability of the OMI prediction. This is to be expected. When two shapes are very similar, ROCS does not have to search far from the inertial alignments, whereas the converse is true for very differently shaped molecules. Another advantage of OMI is that it gives a spatial alignment, a point that will be expanded upon below.

Next we looked at each method in terms of its virtual screening performance. The DUD data set is as close to a standard as the field currently possesses, even though the issue of the congeneric nature of the active classes and the appropriateness of its decoy set is well-known. Figure 30 illustrates the AUC values for the recall of actives using the standard DUD target, actives and decoys, averaged over 40 systems. As

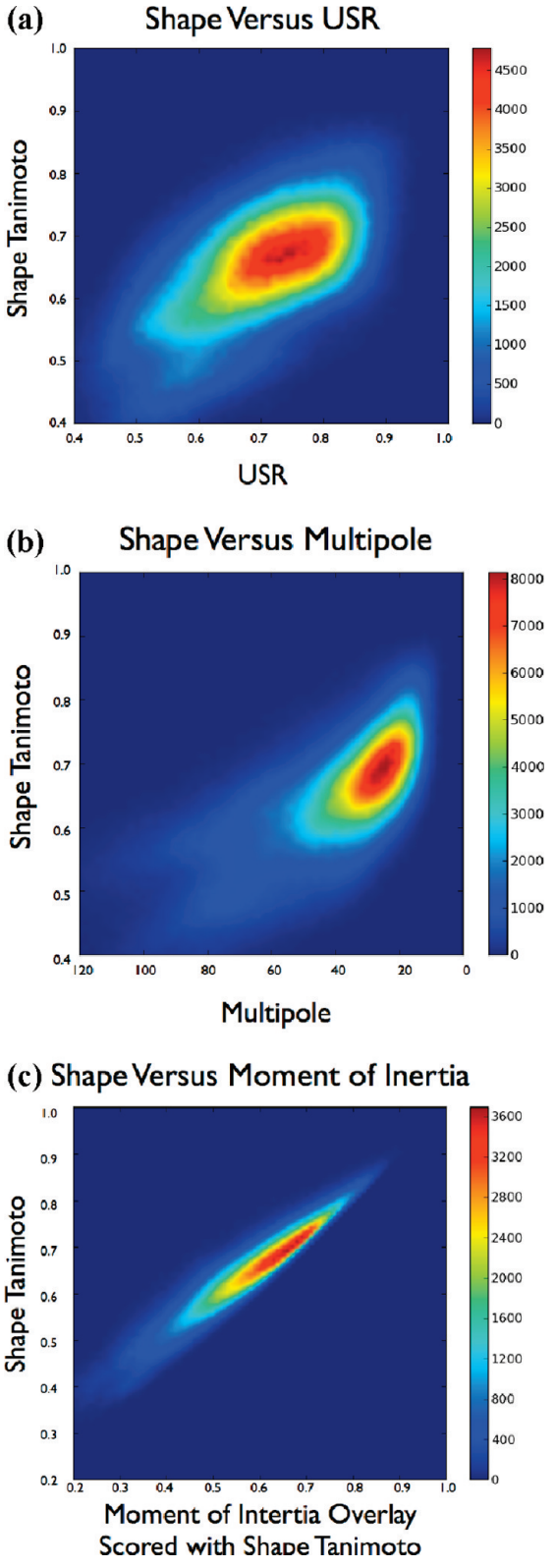

Figure 28. Contour plots of Shape Tanimoto, as calculated by ROCS, versus three approximate methods, USR, steric multipoles, and shape score from the best alignment by moments of inertia.

can be seen, each approximate method has an average AUC approximately the same as that from just using the number of heteroatoms as a descriptor, a truly "blunt" measure and much worse than full ROCS that includes both shape and color descriptions. A positive view of this result is that the

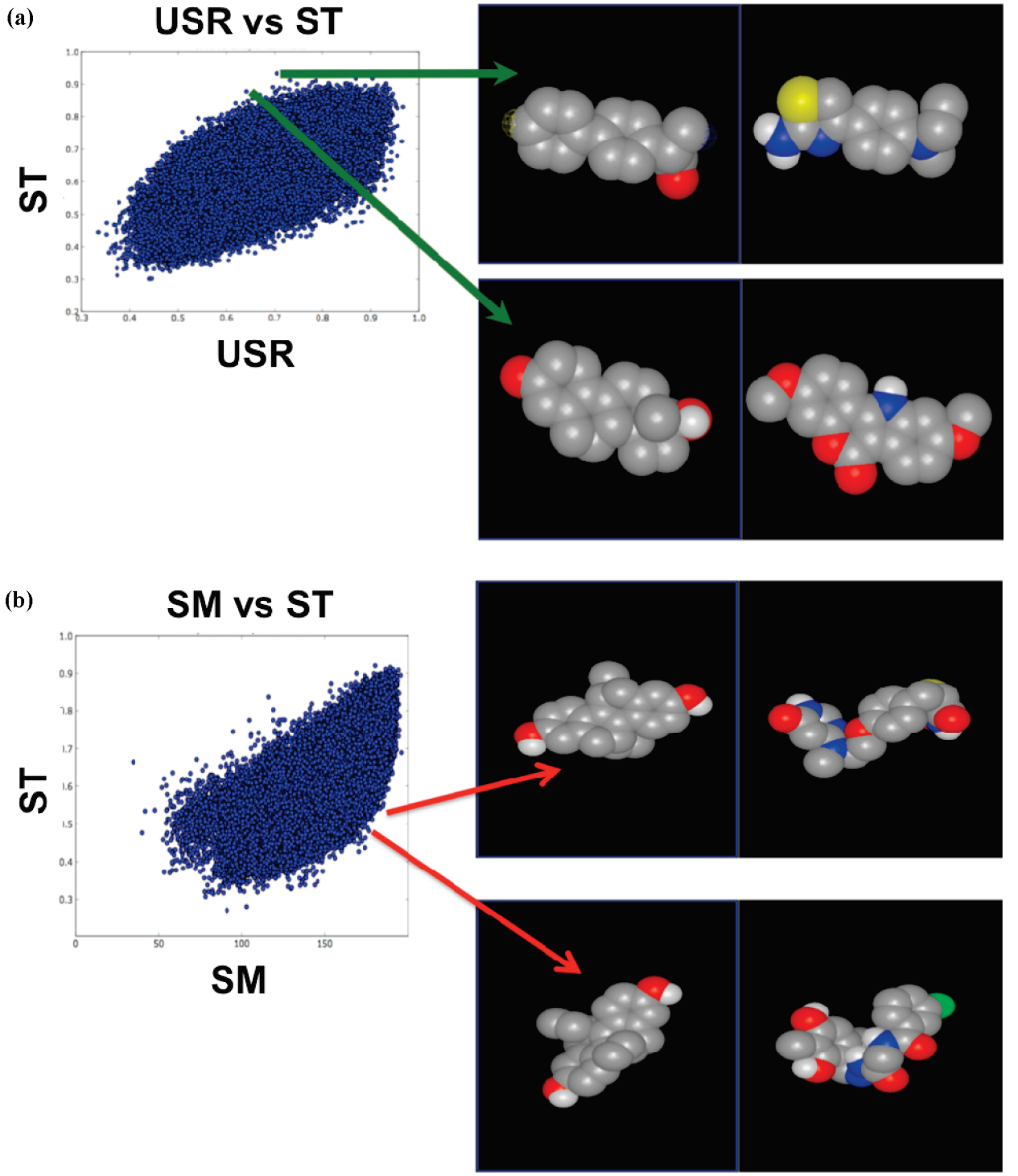

Figure 29. Examples of false negatives pairs from USR, illustrating its fragility as a shape measure, and false positives pairs from steric multipoles, illustrating its bluntness as a measure of shape similarity.

approximate methods are no worse than using the full description of shape. And yet if their correlation to shape is so poor, as we have shown above, this implies that some other feature is inadvertently being encoded that has some very slight value in discerning actives from decoys. These may be entirely artificial, i.e., because although DUD decoys are meant to be property matched, there may still be some less obvious indicators that say "decoy" to some descriptors.

Clearly pure shape does not do very well in this retrospective study. Does this mean that coarse descriptors are as useful as shape? We would claim not because a shape match also includes an alignment in space. This allows the evaluation of 3D chemical similarity that greatly surpasses the other methods considered here, as illustrated in Figure 30. This figure shows what happens if a color score is added to the inertial alignments of the OMI method. This greatly improved performance, at almost no additional cost, illustrates the advantage of methods that are not pure reductions to descriptor sets but rely on the information that comes from a meaningful alignment. In other words, methods that are rotationally invariant may be fast but they also miss the point, or should we say the 'volume'?

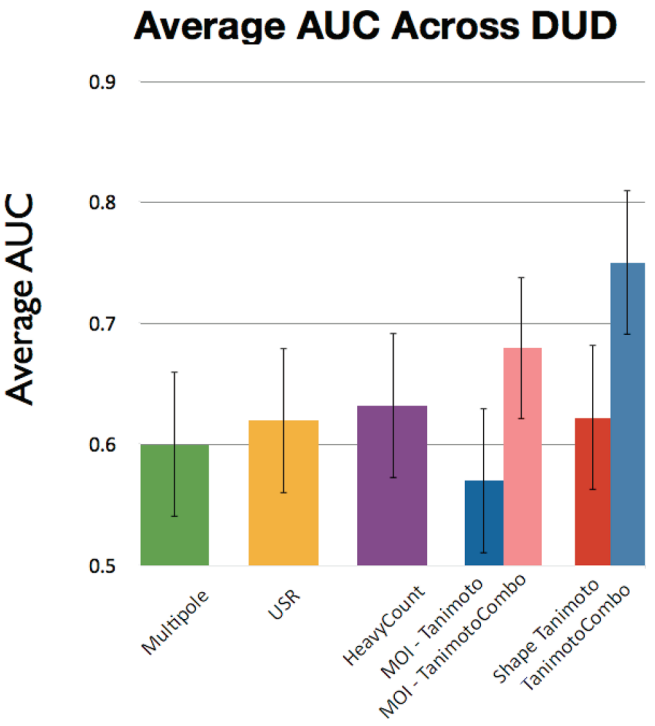

Figure 30. Virtual screening performance on the DUD data set of shape multipoles, USR, the number of heteroatoms ("HeavyCount"), a shape score from the best moments of inertia alignment with and without color, a shape score from the best ROCS alignment with and without color.

Finally, we consider whether approximate methods might act as prescreen tests for more diligence by more CPU costly methods. In the early days of human fingerprint comparison, approximate methods would routinely rely on expert analysis of a set of potential matches, and to this day there is still an element of human intervention after computations are finished. The issue here, then, is the degree to which false positives and false negatives can be tolerated. The former mean the additional testing of unproductive molecules; the latter means the loss of matches entirely. As false negatives cannot be recovered, we suggest erring on the side of methods with more false positives than negative behavior, i.e., blunt rather than fragile. This profile fits the SM profile well; however, if one sets a reasonable threshold for "misses" of 10%, the false positive rate is fairly overwhelming. At this error rate only a 4-fold reduction in conformers is achieved. The problem is just that approximate methods are too noisy. Either too much is lost through fragility or too much is let through by bluntness. In a Pareto-sense, the OMI approach with added color seems to have most to offer. It is fast and captures more of shape than other approximate methods, and the addition of proximity matching of chemistry to its alignments is invaluable.

The problem of approximate shape matching has been considered in other fields, such as robotic vision. Here the problem is the identification of objects against an internal store of known structures, a problem more diverse and difficult than molecular recognition. The field has created many approximate methods, some similar to those considered here, but these are, in general, considered lacking. In his Ph.D. thesis, Kazhdan[106] suggests this is because shape similarity has a well-known metric property; i.e., the distance between two shapes has an identifiable mathematical structure. Methods that do not approximate this metric inevitably fail to capture the essence of shape, whether through fragility or bluntness; they hit frustratingly low glass-ceilings of performance that cannot be breached. Work on shape comparison of molecules recapitulates these observations. They are certainly fast, but they are not good, and even the apparent speed has to be considered in the context of the time taken for the construction of ancillary information, for instance, the necessary molecular conformations. Only when this cost can be amortized over many searches are approximate methods actually faster than thorough shape comparison. In summary, it seems uncertain as to whether approximate methods are good enough to be considered useful.

Summary

The eight contributions here provide ample evidence that shape as a volume or as a surface is a vibrant and useful concept when applied to drug discovery. It provides a reliable scaffold for "decoration" with chemical intuition (or bias) for virtual screening and lead optimization but also has its unadorned uses, as in library design, ligand fitting, pose prediction, or active site description. Computing power has facilitated this evolution by allowing shape to be handled precisely without the need to reduce down to point descriptors or approximate metrics, and the diversity of resultant applications argues for this being an important step forward. Certainly, it is encouraging that as computation has enabled our intuition, molecular shape has consistently surprised us in its usefulness and adaptability.

The first Aurelius question, "What is the essence of a thing?", seems well answered, however, the third, "What do molecules do?", only partly so. Are the topics covered here exhaustive, or is there more to come? To date, there has been little published on the use of the volumetric definition of shape described here as a QSAR variable, for instance, in the prediction or classification of activity, although other shape definitions have been successful applied, for instance, as embodied in the Compass program described above in "Shape from Surfaces". Crystal packing is a phenomenon much desired to be understood. Although powerful models have been applied to the problem,[107] to what degree is this dominated purely by the shape of a molecule? The shape comparison described here is typically of a global nature, and yet some importance must surely be placed on partial shape matching, just as the substructure matching of chemical graphs has proved useful. The approach of using surfaces, as described here, offers some flavor of this, as does the use of metrics that penalize volume mismatch less than the Tanimoto, e.g., Tversky measures. As yet, there is little to go on as to how useful a paradigm this will be because there is less software and fewer concrete results. Finally, the distance between molecular shapes, or between any shapes defined as volumes or surfaces, is a metric property in the mathematical sense of the word. As yet, there has been little, if any, application of this observation. We cannot know what new application to the design and discovery of pharmaceuticals may yet arise from the simple concept of molecular shape, but it is fair to say that the progress so far is impressive.

Acknowledgment. The authors thank Geoff Skillman, Tamsin Mansley, and Marti Head for their help in the construction of this Perspective.

Biographies

Anthony Nicholls received his Ph.D. in Biophysics from Florida State University in 1988. He worked as a Postdoctoral Fellow and later Research Associate with Barry Honig at Columbia University for 7 years, developing the programs DelPhi and GRASP. In 1997 he founded OpenEye Scientific Software and remains the CEO and President of that company. He wrote the introduction "Shape and Medicinal Chemistry" and the "Summary" and contributed to the section "Approximate Shape Methods".

Georgia B. McGaughey received her Ph.D. from the University of Georgia in 1995 and subsequently worked as a Postdoctoral Fellow at Colorado State University examining π-stacking interactions in proteins. In 1997 she went to work at Wyeth Pharmaceuticals focusing on ligand optimization for GPCR receptors. In 1999 she joined Merck Research Laboratories as a member of the Chemistry Modeling and Informatics group. She is currently working in the research areas of lead identification and optimization with focus on neuroscience and antiviral programs. She has authored or coauthored more than 40 scientific publications, numerous of which are focused on virtual screening. She contributed to the section "Shape and Virtual Screening".

Robert P. Sheridan received his Ph.D. in Biochemistry from Princeton University in 1979, working on the application of molecular orbital calculations to enzyme mechanisms. He worked as a NMR spectroscopist at the Fox Chase Cancer Center from 1979 to 1981, after which he did a second postdoc at the Rutgers University Chemistry Department doing molecular dynamics. He first entered the pharmaceutical industry in 1983 when he joined Lederle Laboratories as a molecular modeler. In 1991 he moved to the molecular modeling group of Merck Research Laboratories, Rahway, NJ. His job is to develop and test new modeling/cheminformatic methods, as well as do applications modeling for ADMET. Virtual screening, database mining, and QSAR are his current specialties. He has about 80 publications. He contributed to "Shape and Virtual Screening".

Andrew C. Good received his D.Phil. in Molecular Similarity in 1993 from Oxford University, U.K., under the tutelage of Prof. Graham Richards. He then studied as a postdoctoral fellow at University of California, San Francisco, CA, working on molecular similarity and docking with Prof. Irwin Kuntz. Since then, he has worked on numerous drug discovery projects as a computational chemist at Rhone-Poulenc Rorer, Glaxo Wellcome, and Bristol-Myers Squibb before taking up his current position as a Scientific Director at Genzyme. He has published extensively in the field of computational chemistry and has numerous drug discovery patents in his name. He contributed to the section "Lead Optimization".

Gregory Warren received his Ph.D. in Biochemistry from Massachusetts Institute of Technology, Cambridge, MA. He worked as Postdoctoral Fellow in the laboratory of Axel Brunger as part of the development team for CNS. He worked for 8 years as a molecular modeler at GlaxoSmithKline Pharmaceuticals before moving to OpenEye Scientific Software, Inc. His responsibilities at OpenEye include structure based design and X-ray crystallography applications. He contributed to the "Protein Crystallography" section.

Magali Mathieu, an engineer from Ecole Centrale Paris, completed her Ph.D. in Biophysics in 1995 at the EMBL, Heidelberg, Germany, in the group of R. Wierenga, on the crystal structure of thiolase. She then worked as a Postdoctoral Fellow, first at the AFMB in Marseille, France, and then in the group of F. Rey (first in the LEBS and then the GV) at Gif-sur-Yvette, France, on the structure of viral proteins. In 1999 she started working as a protein crystallographer for Aventis Pharma, now Sanofi-Aventis. She is currently head of their crystallographic research team of Vitry, France, where her group determines the structures of proteins of pharmaceutical interest in complex with potential drugs. She contributed to the "Protein Crystallography" section.

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. Steve's specialty is in biological structure determination by X-ray crystallography. 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. Steve contributed to the section entitled "Pose Prediction".

Scott P. Brown, Ph.D., is an Associate Research Investigator Chemist in Global Pharmaceutical Research and Development at Abbott Laboratories. Scott joined Abbott in September 2003 after completing a postdoc at University of California, Berkeley, CA, in which he developed sampling and model-simulation strategies for improving our understanding of protein folding kinetics and thermodynamics. Scott currently works on a variety of pharmaceutically relevant research projects at Abbott ranging from cheminformatics to molecular modeling. Scott contributed to the section entitled "Pose Prediction".

J. Andrew Grant received B.Sc. and Ph.D. degrees from the University of Sheffield, U.K., followed by postdoctoral studies at Cornell University, Ithaca, NY. He subsequently joined the Computational Chemistry group at AstraZeneca, where he has collaboratively worked on problems connected with molecular shape and electrostatics. He contributed to the section "Library Design".

James A. Haigh has a degree in Chemistry and a Ph.D. in Physical Chemistry from the University of Southampton, U.K. After postdoctoral work in polymer physics with Leo Mandelkern at Florida State University he has been working on shape-based methods as part of a collaboration between the University of Sheffield (U.K.), AstraZeneca, and OpenEye Scientific Software. He contributed to the section "Library Design".

Neysa Nevins received her Ph.D. from the University of Georgia in 1994 parametrizing the MM4 force field and was a postdoctoral fellow at Emory University from 1995 to 1999. She has been working as a computational chemist at GlaxoSmithKline since 2001. She enjoys the opportunity to collaborate with other chemists, biologists, and biochemists working toward developing candidate drug molecules and is aligned with projects in the oncology and respiratory therapeutic areas. Current interests include assessing protein target druggability and improvements in docking and scoring. Before entering industry, she taught physical chemistry and developed a molecular science course for nonscience majors at Elizabethtown College, Elizabethtown, PA. She contributed the section "Binding Site Shape".

Ajay N. Jain earned his Ph.D. in Computer Science in 1991 from Carnegie Mellon University, Pittsburgh, PA. He spent a number of years developing computational methods for military target recognition, but he transitioned from defense applications to drug discovery in a series of biopharmaceutical startup companies, beginning with Arris Pharmaceutical in 1992. After several years in Bay Area biotechnology, including founding BioPharmics LLC, Dr. Jain joined University of California, San Francisco, CA, in 1999, where he was appointed Professor of Bioengineering and Therapeutic Sciences. Shape-based computational methods for drug design form the core of his research approach, which has produced programs such as Compass, Hammerhead, IcePick, and the Surflex family of programs for molecular docking, similarity, and ligand-based activity prediction. Dr. Jain authored the "Shape from Surfaces" section.

Brian Kelley graduated from Cornell University, Ithaca, NY, with a B.A. in Physics and an M.Eng. in Electrical Engineering. He then spent several years at Los Alamos National Laboratory

working on the AMISS project (Advanced Material Information and Security System) as a senior software developer. Several members of this team spun off BioReason Inc., dedicated to organizing and presenting molecular information to medicinal chemists. Later he joined the Whitehead Institute in Cambridge, MA, and wrote "SLIMS" to handle HTS data and was principle author of PathBLAST to help predict druggable protein−protein interactions in *H. pylori*. He has worked at OpenEye since 2004 and contributed to the sections "Pose Prediction" and "Approximate Shape Methods".

Supporting Information Available: Initial 2D schme of ex20, results from Afitt, and data and refinement statistics for ex20. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Augustus, M. A. A. "Meditations", AD167, Chapter 8, Verse 10. Trans. Long, G. "Thoughts of Marcus Aurelius Antoninus", 1862.
(2) Levi, P. *Other People's Trades*; Summit Books: New York, NY, 1989; ISBN 0-671-61149-6.
(3) Anderson, E.; Veith, G. D.; Weininger, D. S*MILES: A Line Notation and Computerized Interpreter for Chemical Structures*; Report No. EPA/600/M-87/021; U.S. EPA, Environmental Research Laboratory—Duluth: Duluth, MN, 1987.
(4) 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.* **1997**, *23*, 3–25.
(5) van Drie, J. H. Pharmacophore discovery: lessons learned. *Curr. Pharm. Des.* **2003**, *9* (20), 1649–1664.
(6) Hahn, M. Three-dimensional shape-based searching of conformationally flexible compounds. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (1), 80–86.
(7) Putta, S.; Beroza, P. Shapes of things: computer modeling of molecular shape in drug discovery. *Curr. Top. Med. Chem.* **2007**, *7* (15), 1514–1524.
(8) Kuhn, T. S. *The Structure of Scientific Revolutions*; University of Chicago Press: Chicago, IL, 1962; ISBN10: 0226458083.
(9) Quine, W. V. *Natural Kinds in Ontological Relativity and Other Essays*; Columbia University Press: New York, 1977.
(10) Clark, D. E. What has virtual screening ever done for drug discovery? *Expert Opin. Drug Discovery* **2008**, *3* (8), 841–851.
(11) Kearsley, S. K.; Smith, G. M. An alternative method for the alignment of molecular structure maximizing electrostatic and steric overlap. *Tetrahedron Comput. Methodol.* **1990**, *3*, 615–633.
(12) Miller, M. D.; Kearsley, S. L.; Underwood, D. J.; Sheridan, R. P. FLOG: a system to select quasi-flexible ligands complementary to a receptor of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 153–174.
(13) Miller, M. D.; Sheridan, R. P.; Kearsley, S. L. SQ: a program for rapidly producing pharmacophorically relevant molecular superpositions. *J. Med. Chem.* **1999**, *42*, 1505–1514.
(14) Hartman, G. D.; Egbertson, M. S.; Halczenko, W.; Laswell, W. L.; Duggan, M. E.; Smith, R. L.; Naylor, A. M.; Manno, P. D.; Lynch, R. J. Non-peptide fibrinogen receptor antagonists. 1. Discovery and design of exosite inhibitors. *J. Med. Chem.* **1992**, *35*, 4649–4642.
(15) McMasters, D. R.; Garcia-Calvo, M.; Maiorov, V.; McCann, M. E.; Meurer, R. D.; Bull, H. G.; Lisnock, J.; Howell, K. L.; Devita, R. J. Spiroimidazolidinone NPC1L1 inhibitors. 1: Discovery by 3D-similarity-based virtual screening. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 2965–2968.
(16) Yang, L.; Guo, L.; Pasternak, A.; Mosley, R.; Rohrer, S.; Birzin, E.; Foor, F.; Cheng, K.; Schaeffer, J.; Patchett, A. A. Spiro[1*H*-indene-1,4′-piperidine] derivatives as potent and selective nonpeptide human somatostatin receptor subtype 2 (sst2) agonists. *J. Med. Chem.* **1998**, *41*, 2175–2179.
(17) Feuston, B. P. M. M. D.; Culberson, J. C.; Nachbar, R. B.; Kearsley, S. K. Comparison of knowledge-based and distance geometry approaches for generation of molecular conformations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 754–763.
(18) Yang, L.; et al. Potent and Selective Non-Peptide Human Somatostatin Receptor Subtype-2 (hSSTR-2) Agonists. *Book of Abstracts*, 216th National Meeting of the American Chemical Society, Boston, MA, Aug 23−27, 1998; American Chemical Society: Washington, DC, 1998.
(19) Walters, W. P.; Stahl, M. T.; Murcko, M. Virtual screening. An overview. *Drug Discovery Today* **1998**, *3* (4), 160–163.
(20) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.
(21) Jain, A. N.; Nicholls, A. Recommendations for evaluations of computational methods for docking and ligand-based modeling. *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 133–139.
(22) Zhang, Q.; Muegge, I. Scaffold hopping through virtual screening using 2D and 3D similarity descriptors: ranking, voting and consensus scoring. *J. Med. Chem.* **2006**, *49*, 1536–1548.
(23) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *Abstracts of Papers*, 234th National Meeting of the American Chemical Society, Boston, MA, Aug 19−23, 2007; American Chemical Society: Washington, DC, 2007.
(24) Sheridan, R. P.; McGaughey, G. B.; Cornell, W. D. Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 257–265.
(25) Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the early recognition problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
(26) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliard, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(27) Sheridan, R. P. Unpublished results.
(28) Sheridan, R. P. Chemical similarity searches: when is complexity justified? *Expert Opin. Drug Discovery* **2007**, *2*, 423–430.
(29) Fan, Y.; Lai, M. H.; Sullivan, K.; Popiolk, M.; Andress, T. H.; Dollings, P.; Pausch, M. H. The identification of neurotensin NTS1 receptor partial agonists through a ligand-based virtual screening approach. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 5789–5791.
(30) Muchmore, S. W.; Souers, A. J.; Akritopoulou-Zanze, I. The use of three-dimensional shape and electrostatic similarity searching in the identification of a melanin-concentrating hormone receptor 1 antagonist. *Chem. Biol. Drug Des.* **2006**, *67*, 174–176.
(31) Rush, T. S.; 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*, 1489–1495.
(32) Stachel, S. J.; Coburn, C. A.; Steele, T. G.; Crouthamel, M. C.; Pietrak, B. L.; Lai, M. T.; Holloway, M. H.; Munshi, S. K.; Graham, S. L.; Vacca, J. P. Conformationally biased P3 amide replacements of beta-secretase inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16* (3), 641–644.
(33) Barrow, J. C.; Stauffer, S. R.; Rittle, K. E.; Ngo, P. L.; Yang, Z.; Selnick, H. G.; Graham, S. L.; Munshi, S.; McGaughey, G. B.; Holloway, M. K.; Simon, A. J.; Price, E. A.; Sankaranarayanan, S.; Colussi, D.; Tugusheva, K.; Lai, M. T. l.; Espeseth, A. S.; Xu, M.; Huang, Q.; Wolfe, A.; Pietrak, B.; Zuck, P.; Levorse, D. A.; Hazuda, D.; Vacca, J. P. Discovery and X-ray crystallographic analysis of a spiropiperidine iminohydrantoin inhibitor of beta-secretase. *J. Med. Chem.* **2008**, *51* (20), 6259–6262.
(34) Lewell, X. Q.; Judd, D.; Watson, S.; Hann, M. RECAP—retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 511–522.
(35) Jahnke, W. Fragment-based approaches. *Comp. Med. Chem. II* **2006**, *3*, 939–957.
(36) Jhoti, H.; Cleasby, A.; Verdonk, M. l.; Williams, G. Fragment-based screening using X-ray crystallography and NMR spectroscopy. *Curr. Opin. Chem. Biol.* **2007**, *11*, 485–493.
(37) Breed. Developed and Distributed by the Chemical Computing Group. http://www.chemcomp.com/software-moe2008.htm (accessed Feb 2009).
(38) Hubbard, R. E.; Davis, B.; Chen, I.; Drysdale, M. J. The SeeDs approach: integrating fragments into drug discovery. *Curr. Top. Med. Chem.* **2007**, *7*, 1568–1581.
(39) Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M. Recore: a fast and versatile method for scaffold hopping based on small molecule crystal structure conformations. *J. Chem. Inf. Model.* **2007**, *47*, 390–399.
(40) Recore. Developed and Distributed by BioSolveIT. http://www.biosolveit.de/ReCore/ (accessed Feb 2009).
(41) ROCS and BROOD. Developed and Distributed by Openeye Scientific Software Inc. www.eyesopen.com/products/applications/brood.html (accessed Feb 2009).

(42) Bergmann, R.; Linusson, A.; Zamora, I. SHOP: scaffold hopping by GRID-based similarity searches. *J. Med. Chem.* **2007**, *50*, 2708–2717.
(43) Fechner, U.; Schneider, G. Flux (2): comparison of molecular mutation and crossover operators for ligand-based de novo design. *J. Chem. Inf. Model.* **2007**, *47*, 656–667.
(44) Mauser, H.; Guba, W. Recent developments in de novo design and scaffold hopping. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 365–374.
(45) Pierce, A. C.; Rao, G.; Bemis, G. W. Generating novel inhibitors through hybridization of known ligands. Application to CDK2, P38, and HIV protease. *J. Med. Chem.* **2004**, *47*, 2768–2775.
(46) Good, A. C. Novel DOCK clique driven 3D similarity database search tools for molecule shape matching and beyond: adding flexibility to the search for ligand kin. *J. Mol. Graphics Modell.* **2007**, *26*, 656–666.
(47) Good, A. C.; Tebben, A.; Claus, B. New Pharmacophore Constrained Gaussian Shape/Electrostatic/Colored Force Field Similarity Searching Tools: Feeding the Synthetic Beast with KIN. *Abstracts of Papers*, 234th National Meeting of the American Chemical Society, Boston, MA, Aug 19−23, 2007; American Chemical Society: Washington, DC, 2007; COMP-249.
(48) Good, A. C.; Hodgkin, E. E.; Richards, W. G. The utilisation of Gaussian functions for the rapid evaluation of molecular similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 188–191.
(49) Good, A. C.; Richards, W. G. Rapid evaluation of shape similarity using Gaussian functions. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 112–116.
(50) DOCK 4. http://dock.compbio.ucsf.edu/Old_Versions/dock4.0_manual.pdf (accessed Feb 2009).
(51) Jones, T. A.; Zou, J. Y.; Cowan, S. W.; Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta. Crystallogr. A* **1991**, *47*, 110–119.
(52) Emsley, P.; Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **2004**, *60*, 2126–2132.
(53) Oldfield, T. J. A Semi-Automated Map Fitting Procedure. *Proceedings of the CCP4 Study Weekend*; Bailey, S., Hubbard, R., Waller, D., Eds.; CLRC Daresbury Laboratory: Warrington, U.K., 1994.
(54) Oldfield, T. J. X-Ligand: an application for the automated addition of flexible ligands into electron density. *Acta Crystallogr. D* **2001**, *57*, 696–705.
(55) PrimeX. http://www.schrodinger.com/ (accessed February 2009).
(56) 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*, 6177–6196.
(57) Wlodek, S.; Skillman, A. G.; Nicholls, A. Automated ligand placement and refinement with a combined force field and shape potential. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006**, *62* (Part 7), 741–749.
(58) Grant, A. J.; Pickup, B. T. A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653–1659.
(59) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74–82.
(60) Wlodek, S.; Skillman, A. G.; Nicholls, A. Automated ligand placement and refinement with a combined force field and shape potential. *Acta Crystallogr. D* **2006**, *62*, 741–749.
(61) Durham, T. B.; Shepherd, T. A. Progress toward the discovery and development of efficacious BACE inhibitors. *Curr. Opin. Drug Discovery Dev.* **2006**, *9*, 776–791.
(62) Hills, I.; Vacca, J. Progress toward a practical Bace-1 inhibitor. *Curr Opin Drug Discovery Dev.* **2007**, *10* (4), 383–391.
(63) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17* (5−6), 490–519.
(64) Halgren, T. A. Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J. Comput. Chem.* **1996**, *17* (5−6), 489–755.
(65) Halgren, T. A. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *J. Comput. Chem.* **1996**, *17* (5−6), 553–586.
(66) Halgren, T. A.; Nachbar, R. B. Merck molecular force field. IV. Conformational energies and geometries for MMFF94. *J. Comput. Chem.* **1996**, *17* (5−6), 587–615.
(67) Blanc, E.; Roversi, P.; Vonrhein, C.; Flensburg, C.; Lea, S. M.; Bricogne, G. Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT. *Acta Crystallogr. D* **2004**, *60*, 2210–2221.
(68) Bricogne, G. Direct phase determination by entropy maximisation and likelihood ranking: status report and perspectives. *Acta Crystallogr. D* **1993**, *49*, 37–60.
(69) Bricogne, G. The Bayesian statistical viewpoint on structure determination: basic concepts and examples. *Methods Enzymol.* **1997**, *276A*, 361–423.
(70) Roversi, P.; Blanc, E.; Vonrhein, C.; Evans, G.; Bricogne, G. Modelling prior distributions of atoms for macromolecular refinement and completion. *Acta Crystallogr. D* **2000**, *56*, 1313–1323.
(71) Baxter, E. W.; Conway, K. A.; Kennis, L.; Bischoff, F.; Mercken, M. H.; Winter, H. L.; Reynolds, C. H.; Tounge, B. A.; Luo, C.; Scott, M. K.; Huang, Y.; Braeken, M.; Pieters, S. M.; Berthelot, D. J.; Masure, S.; Bruinzeel, W. D.; Jordan, A. D.; Parker, M. H.; Boyd, R. E.; Qu, J.; Alexander, R. S.; Brenneman, D. E.; Reitz, A. B. 2-Amino-3,4-dihydroquinazolines as inhibitors of BACE-1 (beta-site APP cleaving enzyme): use of structure based design to convert a micromolar hit into a nanomolar lead. *J. Med. Chem.* **2007**, *50* (18), 4261–4264.
(72) Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem.* **2004**, *47* (10), 2499–2510.
(73) Barillari, C.; Marcou, G.; Rognan, D. Hot-spots-guided receptor-based pharmacophores (HS-Pharm): a knowledge-based approach to identify ligand-anchoring atoms in protein cavities and prioritize structure-based pharmacophores. *J. Chem. Inf. Model.* **2008**, *48* (7), 1396–1410.
(74) Wolber, G.; Seidel, T.; Bendix, F.; Langer, T. Molecule-pharmacophore superpositioning and pattern matching in computational drug design. *Drug Discovery Today* **2008**, *13* (1−2), 23–29.
(75) 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.
(76) Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **2002**, *16* (7), 521–533.
(77) Stahl, M.; Mauser, H.; Tsui, M.; Taylor, N. R. A robust clustering method for chemical structures. *J. Med. Chem.* **2005**, *48* (13), 4358–4366.
(78) Marialke, J.; Korner, R.; Tietze, S.; Apostolakis, J. Graph-based molecular alignment (GMA). *J. Chem. Inf. Model.* **2007**, *47* (2), 591–601.
(79) Bostrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graphics Modell.* **2003**, *21* (5), 449–462.
(80) Sutherland, J. J.; Nandigam, R. K.; Erickson, J. A.; Vieth, M. Lessons in molecular recognition. 2. Assessing and improving cross-docking accuracy. *J. Chem. Inf. Model.* **2007**, *47* (6), 2293–302.
(81) Bostrom, J. Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. *J. Comput.-Aided Mol. Des.* **2001**, *15* (12), 1137–1152.
(82) Hajduk, P. J. Fragment-based drug design: how big is too big? *J. Med. Chem.* **2006**, *49* (24), 6972–6976.
(83) Lowe, D. Column: In the pipeline. *Chem. World* **2008**, *5* (8), 18.
(84) Haigh, J. A.; Pickup, B. T.; Grant, J. A.; Nicholls, A. Small molecule shape-fingerprints. *J. Chem. Inf. Model.* **2005**, *45* (3), 673–684.
(85) Hahn, M. M.; Leach, A. R.; Harper, G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 856–864.
(86) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The design of leadlike combinatorial libraries. *Angew. Chem., Int. Ed.* **1999**, *38* (24), 3743–3748.
(87) 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* (5), 1308–1315.
(88) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893.
(89) Lipkus, A. H.; Yuan, Q.; Lucas, K. A.; Funk, S. A.; Bartelt, W. F., 3rd; Schenck, R. J.; Trippe, A. J. Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry. *J. Org. Chem.* **2008**, *73* (12), 4443–4451.
(90) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68* (1), 76–90.
(91) Wells, J. A.; McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein−protein interfaces. *Nature* **2007**, *450* (7172), 1001–1019.

(92) Fry, D. C. Protein−protein interactions as targets for small molecule drug discovery. *Biopolymers* **2006**, *84* (6), 535–552.
(93) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11* (5), 425–445.
(94) Murray, C. W.; Auton, T. R.; Eldridge, M. D. Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand−receptor binding affinities and the use of Bayesian regression to improve the quality of the model. *J. Comput.-Aided Mol. Des.* **1998**, *12* (5), 503–519.
(95) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789–6801.
(96) Jain, A. N.; Dietterich, T. G.; Lathrop, R. H.; Chapman, D.; Critchlow, R. E., Jr.; Bauer, B. E.; Webster, T. A.; Lozano-Perez, T. A shape-based machine learning tool for drug design. *J. Comput.-Aided Mol. Des.* **1994**, *8* (6), 635–652.
(97) Jain, A. N.; Harris, N. L.; Park, J. Y. Quantitative binding site model generation: compass applied to multiple chemotypes targeting the 5-HT1A receptor. *J. Med. Chem.* **1995**, *38* (8), 1295–1308.
(98) Jain, A. N.; Koile, K.; Chapman, D. Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994**, *37* (15), 2315–2327.
(99) Dietterich, T. G.; Lathrop, R. H.; Lozano-Perez, T. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* **1997**, *89* (1−2), 31–71.
(100) Jain, A. N. Morphological similarity: a 3D molecular similarity method correlated with protein−ligand recognition. *J. Comput.-Aided Mol. Des.* **2000**, *14* (2), 199–213.
(101) Jain, A. N. Ligand-based structural hypotheses for virtual screening. *J. Med. Chem.* **2004**, *47* (4), 947–961.
(102) Cleves, A. E.; Jain, A. N. Effects of inductive bias on computational evaluations of ligand-based modeling and on drug discovery. *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 147–159.
(103) Ballester, P. J.; Finn, P. W.; Richards, W. G. Ultrafast shape recognition: evaluating a new ligand-based virtual screening technology. *J. Mol. Graphics Modell.* **2009**, *27* (7), 836–845.
(104) Ballester, P. J.; Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* **2007**, *28* (10), 1711–1723.
(105) Grant, A. J.; Pickup, B. T. Gaussian Shape Methods. In *Computer Simulations of Biomolecular Systems*; Van Gunsteren, W., Weiner, P., Wilkinson, A. W., Eds.; Kluwer/Escom: Dordrecht, The Netherlands, 1998; pp 150−176.
(106) Kazhdan, M. M. Shape representations and Algorithms for 3D Model Retrieval. http://www.cs.princeton.edu/gfx/pubs/_2004_SRA/thesis.pdf (accessed May 9, 2004).
(107) Day, G. M.; Motherwell, W. D.; Ammon, H. L.; Boerrigter, S. X.; Della Valle, R. G.; Venuti, E.; Dzyabchenko, A.; Dunitz, J. D.; Schweizer, B.; van Eijck, B. P.; Erk, P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W.; Leusen, F. J.; Liang, C.; Pantelides, C. C.; Karamertzanis, P. G.; Price, S. L.; Lewis, T. C.; Nowell, H.; Torrisi, A.; Scheraga, H. A.; Arnautova, Y. A.; Schmidt, M. U.; Verwer, P. A third blind test of crystal structure prediction. *Acta Crystallogr. B* **2005**, *61* (Part 5), 511–527.
(108) Lauri, G.; Bartlett, P. A. CAVEAT: a program to facilitate the design of organic molecules. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 51–66.
(109) Groebke Zbinden, K.; Banner, D. W.; Hilpert, K.; Himber, J.; Lavé, T.; Riederer, M. A.; Stahl, M.; Tschopp, T. B.; Obst-Sander, U. Dose-dependent antithrombotic activity of an orally active tissue factor/factor VIIa inhibitor without concomitant enhancement of bleeding propensity. *Bioorg. Med. Chem. Lett.* **2006**, *14*, 5357–5369.
(110) Bostrom, J. Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. *J. Comput.-Aided Mol. Des.* **2001**, *15* (12), 1137–1152.
(111) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A fast method of molecular shape comparison. A simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17* (14), 1653–1666.