J Comput Aided Mol Des (2008) 22:193–199
DOI 10.1007/s10822-008-9189-4

Community benchmarks for virtual screening

John J. Irwin

Received: 28 November 2007 / Accepted: 30 January 2008 / Published online: 14 February 2008

Abstract Ligand enrichment among top-ranking hits is a key metric of virtual screening. To avoid bias, decoys should resemble ligands physically, so that enrichment is not attributable to simple differences of gross features. We therefore created a directory of useful decoys (DUD) by selecting decoys that resembled annotated ligands physically but not topologically to benchmark docking performance. DUD has 2950 annotated ligands and 95,316 property-matched decoys for 40 targets. It is by far the largest and most comprehensive public data set for benchmarking virtual screening programs that I am aware of. This paper outlines several ways that DUD can be improved to provide better telemetry to investigators seeking to understand both the strengths and the weaknesses of current docking methods. I also highlight several pitfalls for the unwary: a risk of over-optimization, questions about chemical space, and the proper scope for using DUD. Careful attention to both the composition of benchmarks and how they are used is essential to avoid being misled by overfitting and bias.

Keywords Virtual screening · Benchmarking · Enrichment · Decoys

Introduction

Virtual screening is the most practical way to leverage structure for ligand discovery, and as a result many docking programs have been developed [1–7]. Virtual screening orients and scores a library of small molecules in the binding site of a protein, ranking the compounds from best to worst. Often, only a few of the top scoring hits selected for experimental testing actually bind. Yet despite a low hit rate and obvious methodological shortcomings, docking has been influential because purchasable or in house compound predictions can be tested rapidly, and even a couple of novel hits can be very useful. Docking can access far more chemistry, much faster, and at far lower cost than using high throughput screening (HTS). There is thus much current interest in improving docking methods.

The ultimate test of virtual screening is the prospective prediction of new ligands. In principle, one would therefore like to evaluate docking by its ability to prospectively predict ligand binding affinities, but this is now beyond the field. In practice, docking performance is evaluated in retrospective calculations using two criteria. *Pose-fidelity*, the ability to reproduce experimentally observed poses within some tolerance limit—often 2 Å rmsd—is essential, and has been well studied [8–18]. Just as important is *enrichment*, the ability to enrich actives from among a database of decoys, where a decoy is a member of the database that does not bind to the target (Fig. 1). Enrichment is essential for docking to be useful for prioritizing libraries, since molecules that do not score well are unlikely to be tested, no matter how beautiful the pose. For enrichment factors to be meaningful in making comparisons between methods, decoy sets must contain a sufficiently high proportion of decoys that will be challenging to each of the methods when compared to active ligands. Property matching of decoys to ligands can help in designing challenging decoy sets. As Verdonk, Rognan, Jain and others have pointed out, if the ligands and decoys differ by physical properties alone then the docking may

J. J. Irwin (✉)
Department of Pharmaceutical Chemistry, University of California San Francisco, P.O. Box 2550, Byers Hall, San Francisco, CA 94158-2330, USA
e-mail: jji@cgl.ucsf.edu

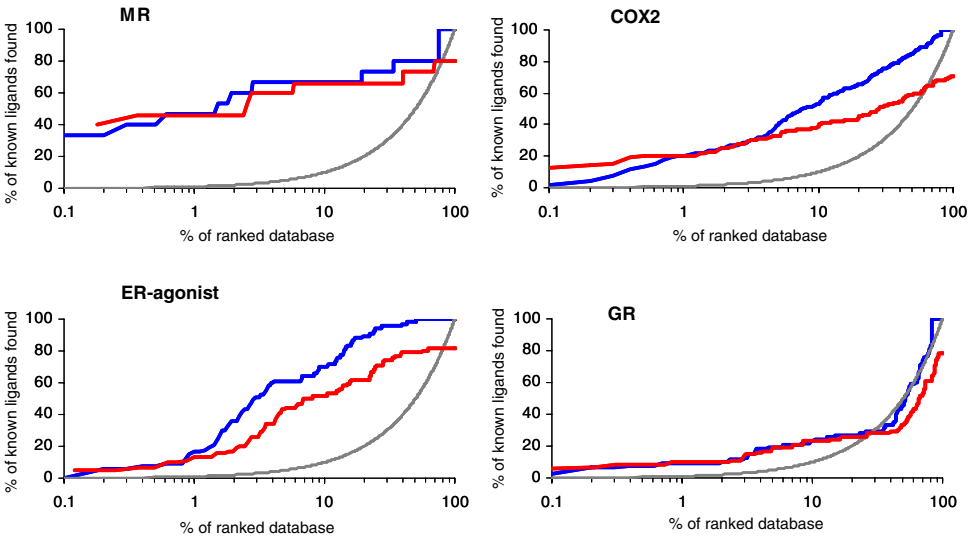

Fig. 1 Enrichment plots showing the enrichment of annotated ligands from a DOCK-ranked database of property-matched (DUD) decoys, comparing the performance of an expert (blue) to a fully automatic program (red) to random performance (grey). Abbreviations. MR, mineralocorticoid receptor; COX2, cyclooxygenase-2; ER-agonist, estrogen receptor agonist; GR, glucocorticoid receptor

separate ligands from decoys, or even worse be *trained* to separate ligands from decoys, by trivial properties such as molecular weight, hydrophobicity, or polarity [13, 19–21]. DUD is an effort to reduce these biases and to give the community a common touchstone by which to evaluate itself. Whereas several other benchmarks have been introduced, in this essay I will restrict my comments to DUD, because I know its weaknesses, and because it has achieved at least partial acceptance as a benchmark.

Three caveats

As useful as virtual screening benchmarks can be, there are three reasons to use them carefully. First and foremost, since programs can and do become trained to meet benchmarks, there is a real worry of overfitting, for instance by selecting the coefficients of a scoring function based on the best retrospective performance. Second, there are important and hard-to-quantify biases in the ligands used in any benchmark. Not only are the annotated actives almost certainly an incomplete index of what will bind, but the pool from which DUD decoys were drawn, ZINC [22], is itself a massively incomplete sampling of chemical space. ZINC is heavily biased towards compounds that are either economical to make (e.g. using combinatorial or parallel strategies with inexpensive intermediates), or are similar to compounds for important targets (e.g. focused kinase libraries, benzodiazepines). Third, DUD is specifically designed to address the weaknesses of docking methods and is not a benchmark for any other purpose. In particular, it is ill suited to evaluate the performance of ligand-based approaches. Failure to pay attention to any of these issues could lead a benchmarking set to do more harm than good.

Overfitting

Everyone trains to benchmarks. It is simple human nature to want to make our programs work better on goals that we set. Indeed, the scoring functions for GOLD, FlexX and Glide, for example, were explicitly parameterized to reproduce crystallographic poses [1, 23, 24] and sometimes even binding affinities. Yet as Ferrari has shown, optimizing for experimental poses alone tends to relax the van der Waals and increase the polar contribution to the scoring function [25]. Alarmingly, as she optimized the scoring function for pose fidelity, enrichment fell. Given the way the GOLD, FlexX and Glide scoring functions were tuned, it is hardly surprising that they do well at reproducing crystallographic poses, particularly those resembling one or more cases in the benchmarking set. One worries, though, about whether the goal of pose fidelity has biased the scoring functions, and perhaps has traded enrichment (which was not a target of scoring function optimization) for pose fidelity. If DUD does become a widespread dataset not only for benchmarking, but also training, there is an additional worry that scoring functions and protocols will be tweaked to maximize enrichment, in the same way they have been optimized to reproduce experimental geometries.

From the narrow view of performance against a benchmark, tuning a scoring function makes sense. Indeed, programs will be compared anyway, so why should developers not tune scoring functions to maximize their performance? One reason, simply stated, is the Kubinyi Paradox, originally formulated for QSAR [26]. As the retrospective prediction of historical (training) data is improved by tweaking the model, there is a tendency for the model to make poorer prospective predictions. The explanation is that the model fits not only the signal but

also the noise, so that prospective predictions contain the model error as well as the experimental error.

Overfitting models to noise is a common problem with well-known solutions. Cross-validation, for instance, in which some data are set aside for testing changes in the model, is routinely and widely used crystallography [27], QSAR [26] and machine learning [28]. But molecular docking scoring function development and docking protocol optimization still do not systematically use cross validation to guard against overfitting. One reason may be that the discrete and non-random nature of protein binding sites and the ligands that bind to them make it unclear what exactly should be set aside. Since inter-target variation in performance in docking methods is so high, outcomes can depend enormously on which datasets are "randomly" selected. By setting aside some fraction of the datasets in DUD for testing by cross validation, overfitting at all stages of the virtual screening protocol could be guarded against (Fig. 2). How to choose the cross validation set, and how many targets to set aside, is unclear. On the one hand, setting aside 10% of targets (4 in the case of DUD) is probably too few to be a meaningful test of improvement, whereas setting aside 50% (20 for DUD) is probably too many, since only half the data would be used for optimization. What is certain is that without some kind of control, the risk of overfitting at some stage of virtual screening benchmarking is likely, and worrying. Other approaches to optimize scoring functions and protocols, such as the use of model systems [30] for example, are possible. Given the importance attached to benchmarks, it is difficult to prevent them from being used to optimize scoring functions. Users, and particularly developers, should be aware of the risks of overfitting, and should take reasonable steps to guard against doing so.

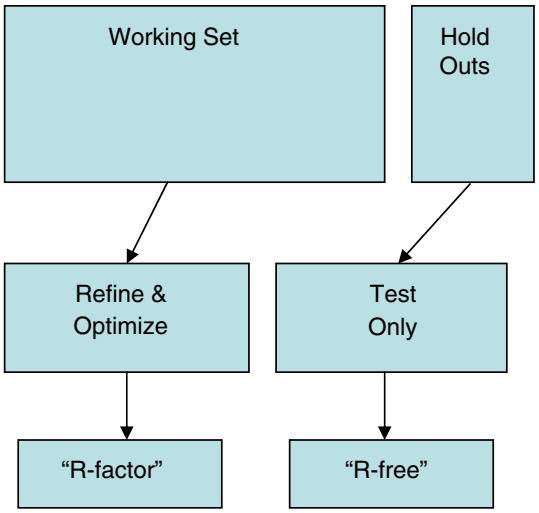

Fig. 2 A proposal to avoid overfitting docking protocols and parameters to benchmarking data. Hold out some fraction (somewhere between 10% and 50%) of any benchmark for testing. Use only the working set for parameter and methodological optimization. Accept changes only if both the "R-factor" and the "R-free" improve. Adapted from Kleywegt [29]. We like his three-tiered approach, separating Refinement, Optimization, and Validation, but have simplified his approach here to make it more likely to be considered for virtual screening

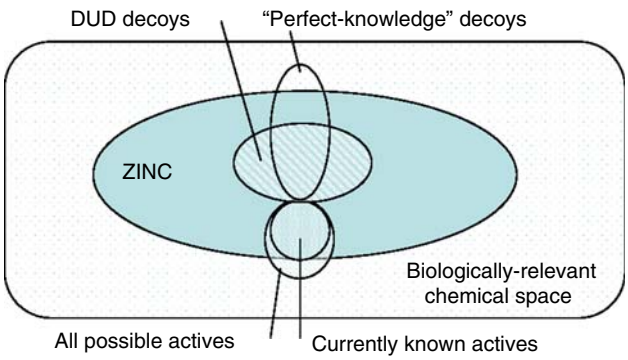

Fig. 3 Schematic of the problem of chemical space and decoys. Decoys are drawn from ZINC, and derived from currently known actives, both of which are biased by incompleteness. (not to scale ;-))

Chemical space

A second reason for caution when using benchmarks such as DUD concerns the sampling of chemical space. Decoys are a proxy for everything that does not bind to the protein, and as such will always be a coarse approximation of this infinite space (Fig. 3). Any selection of a few thousand or even tens of thousands of decoys will always be a massively incomplete proxy. Property-matched decoys such as those in DUD are more useful proxies than randomly chosen molecules because they explicitly challenge docking to differentiate molecules based on specific chemical recognition, not differences in physical properties. Two distinct biases concerning how decoys are selected are incomplete knowledge of the ligands (unknown knowns), and an incomplete pool from which decoys are drawn (unknown unknowns).

Decoys are selected based on physical properties of the annotated ligands, which are almost certainly incomplete, perhaps massively so. Pragmatically, we must accept that there will always be unknown knowns. But besides being aware of the incompleteness of our knowledge, we see how important it is to capture as many actives as possible from the literature, particularly ones that do not resemble those already in DUD. Each new, different annotated active is not only a new challenge for the benchmark, but brings with it a posse of property-matched decoys that would otherwise have been omitted.

DUD decoys are drawn from ZINC, which is an incomplete and biased sampling of biologically relevant

chemical space. For example, when we went to look for decoys for GART, we noticed that there were very few compounds with −2 charge. For HIVPR, there were few large molecules at the high molecular weight of HIVPR ligands, since ZINC contained few molecules larger than 400 Daltons at that time. A pragmatic, if sobering, way to see how incomplete ZINC is to compare it to all possible molecules, which can be done if we restrict ourselves to 11 atoms using only C, N, O and F with Reymond's enumeration of 26.4 million molecules [31]. ZINC 2008, with over 10 million commercially available compounds, only has 43,825 that match Reymond's criteria, i.e. 0.17% of chemical space, even for these very small molecules. Worse, the under sampling of chemical space rises sharply with molecular weight [32], so 0.17 % sampling is a gross overestimate.

The strategy of seeking property-matched decoys can tell us when we need to go looking for more. One tactic to ensure a broader and more challenging set of decoys is to demand that decoys differ from each other by at least some minimum amount. Here, we use the Tanimoto coefficient (Tc) to compare standard Daylight chemical (topological) fingerprints [33]. For example, we could add the constraint that all decoys for a particular target must differ not only from all actives by a Tc of 0.6, making them unlikely to bind, but also from each other by a Tc of 0.9, which ensures they are not trivially similar. This constraint serves two purposes. First, it enforces a minimum level of sampling of chemical space via constrained dissimilarity of the ligands and avoids redundancy in the decoys. More importantly, if we are unable to find enough decoys in ZINC that match the properties of annotated compounds and yet differ from each other, it says that our source database is missing an important region of physico-chemical space with similar properties to annotated actives. With too few property-matched decoys for our ligands, misleading enrichment can be expected regardless of how decoys are chosen. If not enough decoys are available in ZINC, new decoys may be generated in silico to ensure non-redundant decoys. When this happens, we also make a note to look for and prioritize new compounds that appear in catalogs for inclusion in ZINC to cover this missing chemistry. In this way, DUD can be used to drive improvements in the coverage of chemical space in ZINC.

Narrow scope

A third warning about DUD is that it is explicitly designed to test docking by presenting compounds with similar physical properties but different chemistries as enforced by a Tanimoto similarity threshold, i.e. DUD explicitly excludes decoys that would be most challenging to a 2D based similarity method. This makes DUD a cinch for 2D methods, by design, and thus a completely inappropriate benchmark for them. A more challenging benchmarking set for 2D methods might be molecular anagrams: molecules that contain the same substructures, thus having many bits in common in their fingerprints, and yet differing enough chemically that the decoys do not resemble the actives. Similarly, DUD is not an appropriate benchmark for 3D molecular similarity methods, since in its current state there are so many targets for which nearly all actives are trivial analogs of a central structure. Thus benchmarking sets for 2D methods or 3D ligand similarity methods should be designed completely differently than one for docking.

Incremental improvements

As important as the above three problems with benchmarks are, they should neither obscure the importance of using benchmarks, nor prevent incremental improvements from being made. The following list of pragmatic proposals for DUD enumerates some ways to make a better benchmark that can benefit the field. Despite the effort that went into creating it, DUD retains important gaps. These include the dominance of enzymes among the target classes, failure to separate ligands for given targets into well-balanced sets, and failure to include important physical properties in the decoy-ligand comparison, such as charge. There are also improvements to be made in molecular representation, keeping track of errata and corrections, and addressing the issue of standardized protocols. We take up each point in turn.

More targets

Of the 40 targets in DUD, 30 are enzymes, and even among these there are important biases. Adding more and different binding sites to the benchmark is an obvious way to cover more of "target space". Comprehensive coverage of every possible binding site situation and pharmaceutical target class [34] is impractical because such a benchmark would be too big and unwieldy; it is also unrealistic for two practical reasons. First, no crystal structure or even good comparative model is available for perhaps half of all known drug targets. Second, for most ligand binding sites in the PDB, few if any ligands are known that could be used as controls. Still, there are important classes for which sufficient data is available today such as the β2-adrenergic receptor [35], a G-protein coupled receptor, ion channels, such as the potassium ion channel KcsA [36], and immunological proteins such as integrins [37]. Ready-to-use

sources of information for this step include the PDB itself [38], Binding MOAD [39], Binding DB [40], and KiBank [41]. DUD could be doubled immediately, and might even be quadrupled with an intensive curatorial effort.

Ligands per target

Large numbers of highly similar ligands for some targets in the first version of DUD could produce unexpectedly good (or poor) enrichments, even with challenging decoys . To prevent bias, I suggest clustering ligands by Tanimoto similarity to pick representatives from among highly similar ligands. This would help to avoid over-weighting what is essentially a single ligand doing well, or poorly, multiple times, and so biasing the results. Indeed, Good makes this case well elsewhere in this issue [42]. Having similar numbers of ligands per target would make comparison among targets more straightforward for methods that depend on the number of ligands. A reasonable goal for the number of ligands per target might be 100: enough that results are not highly sensitive to a few artifacts, yet not so many that the set (and the corresponding decoys) becomes unwieldy.

Better decoys

Despite the effort that went into matching physical properties among ligands and decoys in DUD, unexpected biases slipped through. Among the most egregious was our failure to insist that every ligand bearing a formal charge was matched by a similar percentage of decoys. Thus, in the current DUD database 41.9% of the ligands are charged, but only 15.2% of the decoys are. This difference is exaggerated in highly charged targets such as thrombin, GART and DHFR, and can lead to weird results when evaluating docking algorithms. For instance, in both GART and DHFR, if one leaves out desolvation altogether—which is highly non-physical—one achieves a better enrichment than if one uses a more correct calculation that includes desolvation. This is simply a matter of decoy bias. Without desolvation, the highly charged ligands will not pay a penalty for their high polarity and will over-dominate the largely neutral decoy molecules. Future benchmarks should have decoys sets that match the formal charge distribution of the ligands.

Molecular representation

Benchmarking datasets are composed of molecules that can be prepared in various ways, for example considering charge, protonation and tautomeric forms. The precise recipe for preparing the molecules is crucial to the success of the benchmark, and should be thoroughly investigated. DUD was prepared using the ZINC protocol [22], which attempts to represent the physiologically relevant form of molecules in a pH-dependent fashion. Aromatic nitrogen bases, for example, often have several pKas near physiological pH, making representation a challenging and ongoing curatorial effort. Since any molecular preparation protocol is likely to be imperfect, it is important that the details of molecular representation be spelled out as part of any benchmark.

Benchmarks evolve

Community benchmarking standards present challenges for long term consistency and availability of data. For example, errors in protonation, tautomerization, or even chirality may be discovered after the benchmark has been published. To address this, we offer both DUD (http://dud.docking.org/) and ZINC (http://zinc.docking.org/) as community resources that will continue to be available on an ongoing basis. We offer our wiki (http://wiki.compbio.ucsf.edu/wiki/) for discussion and annotation of all matters pertaining to benchmarking and databases as a community resource, including the annotation of errors in DUD and ZINC. For example, when a chirality error is detected by an investigator working with DUD, he may directly annotate the problem on our wiki, without contacting us. This annotation becomes public and visible for all future visitors to the DUD website. When a new version is released, all annotations on the Wiki will be considered for incorporation into the new release. As a rule, we will not change the database once it is released, to provide consistency, but will apply corrections to future releases.

Beyond benchmarking databases

To compare docking programs and protocols, benchmarking databases are an important but incomplete standard. For example, even the same program on the same target using the same database can yield different top scoring hits depending on binding site preparation and docking parameters. A standardized docking protocol would help create a reproducible standard, at least within the context of any particular docking program. A standard docking protocol, implemented as an executable script in combination with DUD, would be a truly reproducible benchmark. An expert might well do better than the automatic script, particularly if additional knowledge of the site is available. But an automated script would help eliminate expert

knowledge as a variable that now plagues comparisons of docking protocols, and could thus help to advance the field.

Summary and outlook

Docking is an important tool for ligand discovery that can yield useful results, but can also be unpredictable. Benchmarking using community standards offers the hope that docking performance can be systematically and reproducibly assessed. Whereas there probably is no such thing as a perfect benchmarking set, many small improvements can and should be made. Users of DUD or other benchmarking sets for virtual screening should keep in mind three caveats. First, there is a natural tendency towards overfitting, and some kind of cross-validation of any benchmark-motivated methodological improvements should be used. Second, the vastness of chemical space and the incompleteness of both current libraries and current sets of annotated actives means that any benchmark will have important, difficult to detect biases, which can have significant impact on results. Third, each methodology requires its own benchmarks designed with the weaknesses of the method in mind. DUD is a benchmarking set specifically for 3D virtual screening.

Acknowledgements Supported by NIH grant GM71896 (to Brian K. Shoichet and J.J.I.). I thank Prof. Brian K. Shoichet for comments and suggestions arising from an ongoing discussion of this topic, and Dr. Peter Kolb, Kristin Coan and Michael Mysinger for reading the manuscript. I thank the reviewers for thoughtful and helpful suggestions.

References

1. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Improved protein-ligand docking using GOLD. Proteins 52:609–623
2. Abagyan RA, Totrov MM, Kuznetsov DA (1994) ICM: a new method for structure modeling and design. J Comput Chem 14:488–506
3. Meng EC, Shoichet BK, Kuntz ID (1992) Automated docking with grid-based energy evaluation. J Comput Chem 13:505–524
4. McGann MR, Almond HR, Nicholls A, Grant JA, Brown FK (2003) Gaussian docking functions. Biopolymers 68:76–90
5. Friesner RA et al (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47:1739–1749
6. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261:470–489
7. Miller MD, Kearsley SK, Underwood DJ, Sheridan RP (1994) FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. J Comput Aided Mol Des 8:153–174
8. Perola E, Walters WP, Charifson PS (2004) A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins 56:235–249
9. Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks CL III (2004) Assessing scoring functions for protein-ligand interactions. J Med Chem 47:3032–3047
10. Kellenberger E, Rodrigo J, Muller P, Rognan D (2004) Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins 57:225–242
11. Kontoyianni M, McClellan LM, Sokol GS (2004) Evaluation of docking performance: comparative data on docking algorithms. J Med Chem 47:558–565
12. Wang R, Lu Y, Fang X, Wang S (2004) An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. J Chem Inf Comput Sci 44:2114–2125
13. Verdonk ML et al (2004) Virtual screening using protein-ligand docking: avoiding artificial enrichment. J Chem Inf Comput Sci 44:793–806
14. Xing L, Hodgkin E, Liu Q, Sedlock D (2004) Evaluation and application of multiple scoring functions for a virtual screening experiment. J Comput Aided Mol Des 18:333–344
15. Onodera K, Satou K, Hirota H (2007) Evaluations of molecular docking programs for virtual screening. J Chem Inf Model 47:1609–1618
16. Zhou Z, Felts AK, Friesner RA, Levy RM (2007) Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets. J Chem Inf Model 47:1599–1608
17. Hartshorn MJ et al (2007) Diverse, high-quality test set for the validation of protein-ligand docking performance. J Med Chem 50:726–741
18. Nissink JW et al (2002) A new test set for validating predictions of protein-ligand interaction. Proteins 49:457–471
19. Kuntz ID, Chen K, Sharp KA, Kollman PA (1999) The maximal affinity of ligands. Proc Natl Acad Sci USA 96:9997–10002
20. Pham TA, Jain AN (2006) Parameter estimation for scoring protein–ligand interactions using negative training data. J Med Chem 49:5856–5868
21. Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases: 1. evaluation of different docking/scoring combinations. J Med Chem 43:4759–4767
22. Irwin JJ, Shoichet BK (2005) ZINC–a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182
23. Halgren TA et al (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 47:1750–1759
24. Gohlke H, Hendlich M, Klebe G (2000) Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol 295:337–356
25. Ferrari AM, Wei BQ, Costantino L, Shoichet BK (2004) Soft docking and multiple receptor conformations in virtual screening. J Med Chem 47:5076–5084
26. van Drie JH (2003) Pharmacophore discovery–lessons learned. Curr Pharm Des 9:1649–1664
27. Brünger A (1992) The free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature 355: 472–474
28. Byvatov E, Schneider G (2003) Support vector machine applications in bioinformatics. Appl Bioinformatics 2:67–77
29. Kleywegt GJ (2007) Separating model optimization and model validation in statistical cross-validation as applied to crystallography. Acta Crystallogr D Biol Crystallogr 63:939–940
30. Graves AP, Brenk R, Shoichet BK (2005) Decoys for docking. J Med Chem 48:3714–3728
31. Fink T, Reymond JL (2007) Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties,

compound classes, and drug discovery. J Chem Inf Model 47: 342–353
32. Hann MM, Oprea TI (2004) Pursuing the leadlikeness concept in pharmaceutical research. Curr Opin Chem Biol 8:255–263
33. James CA (2007) Daylight Theory Manual 4.93
34. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? Nat Rev Drug Discov 5:993–996
35. Cherezov V et al (2007) High-Resolution Crystal Structure of an Engineered Human {beta}2-Adrenergic G Protein Coupled Receptor. Science 366
36. Yohannan S, Hu Y, Zhou Y (2007) Crystallographic study of the tetrabutylammonium block to the KcsA K+ channel. J Mol Biol 366:806–814
37. Xiong JP et al (2002) Crystal structure of the extracellular segment of integrin alpha Vbeta3 in complex with an Arg-Gly-Asp ligand. Science 296:151–155
38. Berman HM et al (2000) The protein data bank. Nucl Acid Res 28:235–242
39. Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P, Nerothin J, Carlson HA (2008) Binding MOAD, a high-quality protein-ligand database. NAR 36:D674–D678
40. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res 35: D198–201
41. Zhang J et al (2004) Development of KiBank, a database supporting structure-based drug design. Comput Biol Chem 28: 401–407
42. Good AC, Oprea TI (2008) Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? J Comput Aided Mol Des this issue, doi:10.1007/s10822-007-9167-2