J Comput Aided Mol Des (2008) 22:133–139
DOI 10.1007/s10822-008-9196-5

Recommendations for evaluation of computational methods

Ajay N. Jain · Anthony Nicholls

Received: 5 February 2008 / Accepted: 7 February 2008 / Published online: 13 March 2008

Abstract The field of computational chemistry, particularly as applied to drug design, has become increasingly important in terms of the practical application of predictive modeling to pharmaceutical research and development. Tools for exploiting protein structures or sets of ligands known to bind particular targets can be used for binding-mode prediction, virtual screening, and prediction of activity. A serious weakness within the field is a lack of standards with respect to quantitative evaluation of methods, data set preparation, and data set sharing. Our goal should be to report new methods or comparative evaluations of methods in a manner that supports decision making for practical applications. Here we propose a modest beginning, with recommendations for requirements on statistical reporting, requirements for data sharing, and best practices for benchmark preparation and usage.

Keywords Docking · Molecular similarity · Benchmarking · Statistical evaluation

Introduction

The field of computational chemistry, particularly as applied to drug design, has become increasingly important in terms of the practical application of predictive modeling to pharmaceutical research and development. Tools for exploiting protein structures or sets of ligands known to bind particular targets can be used for binding-mode prediction, virtual screening, and quantitative prediction of activity. A serious weakness within the field is a lack of standards with respect to statistical evaluation of methods, data set preparation, and data set sharing. Our goal should be to report new methods or comparative evaluations of methods in a manner that supports decision making for practical applications. In this editorial, we propose a modest beginning, with recommendations for requirements on statistical reporting, requirements for data sharing, and best practices for benchmark preparation and usage.

There are two fundamental premises in making such a proposal. First, we must believe that the goal of reporting new methods or evaluations of existing methods is to communicate the likely real-world performance of the methods in practical application to the problems they are intended to solve. Ideally, the specific relationship between methodological advances and performance benefits will be clear in such reports. Second, we must understand that the utility of the methods of broad utility in pharmaceutical research application are *predicting* things that are *not* known at the time that the methods are applied. While this seems elementary, a substantial proportion of recent reports within the field run afoul of this observation in both subtle and unsubtle ways. Rejection of the first premise can reduce scientific reports to advertisements. Rejection (or just misunderstanding) the second premise can distort any conclusions as to practical utility.

This special issue of the Journal of Computer-Aided Molecular Design includes eleven papers, each of which makes a detailed study of at least one aspect of methodological evaluation [1–11]. The papers collected within this issue make the detailed case for the recommendations that

A. N. Jain (✉)
University of California San Francisco, Box 0128,
San Francisco, CA 94143-0128, USA
e-mail: ajain@jainlab.org

A. Nicholls (✉)
OpenEye Scientific Software, 9 Bisbee Court,
Suite D, Santa Fe, NM 87508, USA
e-mail: anthony@eyesopen.com

follow; the recommendations are intended to provide guidance to editorial boards and reviewers of work submitted for publication in our field. In surveying the eleven papers, we feel there are three main areas of concern: data sharing, preparation of datasets, and reporting of results. Concerns within each area relate to three main subfields of molecule modeling, i.e. virtual screening, pose prediction, and affinity estimation, and to whether protein structural information is used or not. We describe the issues in each area and then present recommendations drawn from the papers herein.

Data sharing

The issues

Reports of new methods or evaluations of existing methods must include a commitment by the authors to make data publicly available except in cases where proprietary considerations prevent sharing. While the details are different across the spectrum of methods, the principle is the same: that sharing data promotes advancement of the field by ensuring study reproducibility and enhancing investigators' ability to directly compare methods. However, the details of this matter a great deal, both for docking methods and for ligand-based methods. Docking will be used to briefly illustrate the problem. Many reports make claims of sharing data by, for example, providing a list of PDB codes for a set of protein–ligand complexes used in evaluating docking accuracy. In a very narrow sense, this might accommodate a notion of sharing. However, this is inadequate for four reasons:

(1) PDB structures do not contain all proton positions for proteins or ligands. Many docking approaches require all atoms, and nearly all require at least the positions of the polar protons. Without the *precise* protein structures used, in a widely used file format, it is not possible to reproduce the results of a report or make comparisons of other methods to those reported [7, 9, 11].
(2) Ligands within PDB structures do not contain bond order information and often do not even contain atom connectivity at all. Lacking this information, it is not possible to know what protonation state or tautomeric state was used to produce a particular result [4, 7–9].
(3) Docking methods have different sensitivities to input ligand geometries, both with respect to absolute pose and with respect to other aspects such as conformational strain and ring conformations. Since docking methods do not search ligand pose space exhaustively, absence of precise input ligand structures produces the same issue of reproduction and comparison as in (1) [4, 7–9].
(4) Different methods of protein structure preparation can yield subtle biases to different types of docking and scoring approaches. Very small changes in heavy atom or proton positions, as come with various relaxation strategies, can yield large changes in the positions of extrema for scoring functions. Provision of coordinates for all atoms allows other investigators to understand and differentiate the effects of methodology from the effects of protein structure preparation [4, 7–9].

Recommendations on data sharing

(1) Authors of reports on methodological advances or methods comparisons must provide *usable* primary data so that their results may be properly replicated and assessed by independent groups. By usable we mean in routinely parsable formats that include all atomic coordinates for proteins and ligands used as input to the methods subject to study. The commitment to share data should be made at the time of manuscript submission.
(2) Exceptions to this should only be made in cases where proprietary data sets are involved for a valid scientific purpose. The defense of such an exception should take the form of a parallel analysis of *publicly available* data in the report in order to show that the proprietary data were required to make the salient points [8].

Preparation of datasets

The issues

As stated earlier, the ultimate goal is *predictions* of things that we *do not already know*. For retrospective studies to be of value, the central issue is the relationship between the information available to a method (the input) to the information to be predicted (the output). If knowledge of the input creeps into the output either actively or passively, nominal test results may overestimate performance. Also, if the relationship between input and output in a test data set does not accurately reflect, in character or difficulty, the operational application of the method to be tested, the nominal reported performance might be unrelated to real world performance. Here, we will briefly frame the issue by discussing the differences between the operational use of methods and the construction of tests to measure and document their effectiveness for both protein structure,

e.g. docking, and ligand-based methods in their areas of application.

Docking

(1) *Pose prediction*. Here the goal is to prove that a method can predict how a ligand may bind, but not whether it can bind. In the operational case, we typically have a protein structure in complex with a ligand (or several such examples). We desire accurate prediction of poses for novel ligands that are potentially quite different from those whose bound structures are known. For method evaluation, the construction of prediction tests varies, but there are two basic forms:

   a. *Cognate docking*. The most common test of pose prediction involves a set of protein structures, each bound to a ligand, and with that ligand being the one to be tested. This represents the easiest form of the problem, since the conformation of the protein contains information pertinent to recovering the correct pose of the ligand. Most commonly, the protein coordinates are used as provided experimentally, with some variation in addition of protons, with the ligand in a randomized starting pose. Examples of information 'leak' include using of the cognate ligand pose as input [7], adding protons to the protein to favor the cognate pose [7, 9], choosing tautomer or charge states based on knowledge of the bound structure [8], and inappropriate use of bridging water molecules [9]. An extreme example would be optimizing the protein–ligand complex under the same scoring function used for docking, and then using this new, non-crystallographic information as the "test" data [7].
   b. *Cross docking*. The less common (but more relevant) formulation employs a protein structure with a bound ligand, but where the ligands to be predicted are different. The issue of similarity between the known ligand and the ligand being tested should be raised, but this is certainly more realistic, since the potential protein rearrangement from the apo form has been partially embedded in the structure but not optimized for each test ligand [7, 8].

(2) *Virtual screening*. Predicting whether a ligand will bind, but not its affinity or its pose. In an operational application, we typically have a protein structure (or several), and we may have a few ligands known to bind a site of interest. The goal is to find novel ligands from some library of compounds. Operationally, we do not have the bound structures of the ligands we are trying to find, nor do we generally have a specific protein structure in which we are guaranteed a hospitable geometry. Many of the same mistakes that can be made with pose prediction can also be made to prefer known ligands over decoys, but there are additional hazards:

   a. The decoys do not form an adequate background [5–8, 10]. One of the frustrations in evaluating a study is how to judge the background against which a method is framed. It is very easy to generate a set of decoys that any method can tell apart from actives, and much more difficult to construct an informative collection.
   b. All the actives are chemically similar [2, 4, 5, 8, 10]. This is more relevant to ligand-based methods, but also applicable to docking because operationally finding chemically similar molecules as being potentially active is of little value in that these will likely be found by other methods.

(3) *Scoring*. Prediction of affinity is the hardest problem for molecular modeling and is as yet unsolved. In the operational case, we typically have multiple protein structures with ligands and may also have a wealth of structure-activity data for multiple ligand series. Frequently the problem here is *accurately* predicting the activity of what may be considered an obvious analog in virtual screening. We do not know the precise bound geometry of the specific ligand whose activity we are predicting.

   a. Affinity prediction tests can be done absent any affinity data on related analogs. However, to date, successful predictions without prior affinity information have been so anecdotal and untransferable that the field seems willing to accept any input of prior structural information. Hence, inclusion of information as to the protein's disposition upon binding that is not available in an operational setting is considered acceptable.
   b. More typically, structural information and the activities of one or more closely related analogs are available. Here there are fairly regular reports of success, if given complete structural information. Chemical similarity is assumed, thus placing this technique in the domain of lead optimization, not lead discovery. As illustrated in at least one of the reports here [3], such methods are not currently successful when properly considered with control computations that include, for example, correlations of affinity with molecular weight.

Ligand-based modeling

(1) *Pose prediction*. This is rarer than the use of ligand information in virtual screening but not operationally uncommon. The goal is to find the alignment of ligands to a protein using one or more known protein–ligand complexes. If the known and predicted ligands are one and the same, then issues from cognate ligand apply, for instance using torsions from the crystal structure, rather than deriving such information. If the known and test ligands are different, then the caveats from cross-docking apply, for instance are the test ligands diverse enough to make this experiment meaningful.

(2) *Virtual screening*. We have some number of ligands known to bind a particular site competitively, or, minimally, a single compound that exhibits a desired activity. The goal is to find novel ligands from some library of compounds. The incremental value of obvious analogs of known ligands is small as such would typically be found from SAR expansion from the known active (and is relevant in the narrow case of expanding hits after, for example, an HTS screen).

a. Quite frequently, test cases are constructed where both the input ligands and testing ligands are all trivial analogs of a common central structure [2, 8, 11]. This stems, in part, from the simple fact that the ligands available for constructing tests are most frequently synthesized as part of a design process in which creating analogs is a useful exercise. However, such test cases do not reflect a key feature of the practical application in lead discovery: ligands that are obvious analogs of existing lead compounds *will not exist* in libraries to be screened for new leads.

b. The relevant test cases are those in which the ligands to be retrieved are not analogs of the input ligands. This is, to a degree, a subjective issue. However, construction of such cases can be done, for example, by choosing input ligands that were discovered long before the test ligands or by choosing input ligands that have substantially different overall biological properties (e.g. side effects) than the test ligands [2].

(3) *Scoring*. Predicting affinities of ligands from the affinity of one or more ligands, whether relative or absolute. In practice, we generally have structure-activity data for multiple ligand series. Frequently the problem here is accurately predicting the activity of what would be considered an obvious analog in virtual screening. We do not generally know the bound geometry of the specific ligand whose activity is to be predicted. This methodological area of QSAR has its own set of relatively well-understood foibles and is not addressed in detail in this issue.

The descriptions of test case construction above involve different degrees of challenge in proportion to the amount of information provided to a method. The problems often encountered in reviewing or reading papers is that methods claim a lower level of information concerning the answers than is actually true. This is seldom intentional, no matter the provocation to believe otherwise, but a reflection of the difficulty in preparing a 'clean' test.

Recommendations on dataset preparation

(1) *Protein structure selection and preparation*.

a. Protein structure selection should take into account more than just the nominal resolution [4, 5, 9]. There are other measures such as coordinate precision that are more appropriate but require the use of structures where an R and R_{free} are reported. In addition, checking to see if density actually exists for the poses being predicted is suggested, although this requires structure factors to have been deposited along with protein coordinates.

b. Protein structure optimization must *not* be done by making use of the known geometry of the ligand that is the subject of a prediction [5, 7]. At most, selection of sensible protonation states, tautomers, and rotamers of ambiguous or underspecified groups should be done one time for each protein structure. Much fuller disclosure of preparation procedures is required than is typically seen.

c. The most relevant tests of methods will employ proteins whose structure was determined with a ligand *other* than the one being predicted or a close analog thereof [8].

d. The number and diversity of protein targets needs to be sufficient to enable to draw statistically robust conclusions [4, 6, 10, 11]. Some typical targets (e.g. HIV protease) are quite atypical [4] and in small datasets may dominate results [10, 11].

(2) *Decoy set construction*. There is clearly a consensus that decoy sets can have a significant impact on results [4–8, 10, 11]. The contributed papers here provide no clear consensus as to what constitutes an acceptable set of decoys, although there are lessons as to what not to do, for instance using molecules that might actually be actives, or have unusual properties compared to known actives. At present, the best suggestions seem to be to make decoys relatively 'drug-like', so as to mimic real, i.e. operational screens. We also recommend the practice of employing multiple decoy sets and including those developed

by other investigators to facilitate study comparison and collation.

(3) *Active ligand set construction*. There is consensus that the degree of "obvious similarity" among actives has important effects, particularly in evaluating ligand-based methods [1, 2, 4, 7], but there is less agreement on how to either measure this or to control for it. Our recommendation is that such effects should be quantified in reports, where possible, by, for example, using 2D similarity methods to provide a baseline for the difficulty of a retrieval task or to provide a numerical characterization of the diversity of active ligand sets [2]. In addition, suggestions are made in this issue to either use only single representatives of a chemical class or to weight each according to its order of discovery [1, 4, 6]. Both ideas seem eminently worth further evaluation.

(4) *Ligand preparation*. All ligands (whether active or decoys) must be prepared using automated procedures that are unbiased and which will not yield systematic differences between populations of molecules that will generate a systematic performance bias [7–9]. For instance, assign protonation states of ligands and decoys by the same protocol, and generate conformations from just connectivity records of both ligands and decoys.

(5) *Parameter tuning*. Many papers in this issue show how the choice of parameters influences the apparent quality of results [3, 4, 9]. There is a dichotomy of opinion on whether "tuned" performance figures are relevant to future application of a method when the correct answer is unknown. Our recommendation is that if one chooses to report tuned performance, one must also report performance using standard parameters.

Even within the constraints outlined above, data set preparation and parameter selection can yield a wide range of results. This is acceptable to illuminate which choices are of most benefit to users of the different methods. However, without strong requirements for data sharing (the subject of the previous section), this benefit will be diluted. Further, without baseline requirements for statistical reporting (the subject of the next section), this diversity will lead to an unacceptable degree of incomparability between different reports.

Reporting results

The issues

The issues surrounding *what* to report are substantially in dispute, and this has led to an alarming inability to compare multiple studies, except in the case where all primary data are available and where one is willing to make an independent analysis. Here there seem to be two schools of thought. The first is that molecular modeling is a special enterprise, distinct and different from other efforts at prediction. As such it is seen as a part of the process to select or invent measures that illustrate a particular point. The second school holds that molecular modeling is in fact similar to many other areas of science and commerce and that by ignoring standard practices in other, more established, fields, we do a disservice to modeling.

(1) *Pose prediction*. The almost universal measure for pose prediction is RMS, i.e. the root-mean-square difference between heavy atom positions seen in crystallographic refinement and predicted by a method, generally corrected to allow for internal symmetries within the ligand in question [8]. What is at issue is the manner in which RMS is reported. The desire, as with enrichment metrics, is for a single value to capture the performance of a method over a collection of test cases. The most commonly reported is the proportion of successes at some particular threshold of RMS (for instance, an arbitrary 2.0 Å RMS), but a number of investigators report average RMS instead. Neither is ideal, but mean RMS is less useful for two reasons. First, it can be skewed by small numbers of poor poses (each with very large RMS) [5]. Second, its magnitude can be directly manipulated by clever choice of poses against which to measure success [5, 7].

(2) *Virtual screening*. In many senses, this is the most disputatious area. The standard measure has been "enrichment" defined to be the ratio of the observed fraction of active compounds in the top few percent of a virtual screen to that expected by random selection. The reason enrichment is so prevalent is that it is synonymous with the purpose of virtual screening: to enable the selection of a subset of compounds with improved chances of drug discovery. However, by nearly all other considerations it is a poor measure. Most regrettable is its dependence on the ratio of actives to inactives, which makes enrichment a property of a method *and* an experimental set-up rather than an intrinsic property of the method [10]. A number of metrics have been proposed, many of which share this clearly undesirable quality [1, 7, 8, 10].

(3) *Affinity estimation*. Ideally the analysis of a prediction of affinity ought to be the simplest of tasks. Given a set of experimental values and a predicted set, merely calculate the average difference. If modeling could actually predict affinities, this might be a reasonable approach. As it is, the best generally hoped for is a

rough correlation between activity and score and even in these cases there are obvious, and not so obvious, pitfalls. However, a rough correlation between activity and score is frequently obtained simply by equating activity with, for example, a monotonic function of molecular weight [3].

(4) *General*. There are other more subtle issues. One is the presentation of results where the answers have fed back to the input (training/test set contamination). This is generally easy to spot and usually means a method is without merit. More subtle errors tend to be where forms for cross-validation are followed (proper separation into training and test systems), but where the true independence of these two sets is never called into question [10]. If the test set is not sufficiently different to the training set then there is no assurance against over-parameterized approaches. Finally, reports that profess to predict affinities seldom provide some reliable estimate of experimental affinity. The practice of combining results from multiple experiments is only acceptable if experimental conditions are similar. Anecdotal stories abound of different labs within the same company failing to be able to reproduce each other's binding affinities, often with difference of an order of magnitude or more. It seems sheer folly to think a test set from truly heterogonous sources can be called reliable.

Recommendations for reporting results

(1) *Pose prediction*. Success rates using multiple RMS thresholds should be reported. At a minimum we recommend 3.0, 2.0, 1.0, and 0.5 Å. In fact, we encourage investigators to report full cumulative histograms of RMS performance for both top scoring and best-predicted poses. This will generally take very little additional space in a report, and it provides much more information to the reader. For example, if there are a large proportion of reported RMS values that appear to have greater precision than the experiment, this is detectable by inspection of the histogram [5, 7]. Statistically it is not impossible in a fair prediction for a measurement to be within, say, 0.1 Å of an experimental measurement that is only accurate to 0.5 Å, but it is unlikely. We also suggest that if an estimate of the precision of the experimental coordinates is available that it must be reported. This, then, provides an excellent bulwark against overfitting to the known results.

(2) *Virtual screening*. Based on multiple reports in this issue, we recommend reporting the area under the curve for ROC plots (AUC) [1, 3, 7–10]. These have for a long time been a standard metric for other fields and for good reasons. The argument against using AUC values to judge methods is that they are global measures, i.e. reflect the performance throughout a ranked list. Thus, the notion of "early enrichment" may not be well characterized by just AUC, particularly when virtual screening methods yield AUC values short of the 0.8–1.0 range. Therefore we make two suggestions. First, enrichment percentages should be reported at the following four values: 0.5%, 1%, 2%, and 5%. Second, that a formulation of enrichment is used that reports the ratio of true positive rates (the Y axis in an ROC plot) to the false positive rates of 0.5%, 1%, 2%, and 5% (found on the X axis in an ROC plot). Thus "enrichment at 1%" becomes the fraction of actives seen along with the top 1% of known decoys (multiplied by 100). This removes the dependence on the ratio of actives and inactives and directly quantifies early enrichment. It also makes standard statistical analysis of error bars much simpler [10].

(3) *Affinity estimation*. First, standard correlation measures must be reported. We recommend Pearson's correlation (due to its intuitive appeal and ubiquity) as well as Kendall's Tau (due to its robustness in cases where Pearson's correlation can yield spurious values). Both are easy to calculate, and errors for both are simple to compute. Second, we recommend that papers claiming a correlation with affinity ought to also present the correlations achieved with simpler measures, to include molecular weight, cLogP, and hydrogen bond donor/acceptor counts [3, 8]. Thirdly, authors must be held responsible for realistic estimates of the accuracy of experimental affinities, in particular when such results are from heterogeneous sources.

(4) *General*. First, if data and dataset preparation are completely disclosed, then the issue of the precise manner of reporting in a paper becomes less vital. Authors may choose to emphasize whatever measures they wish but interested readers should be able to construct alternate measures. Secondly, the most lamentable aspect of reporting in our field is the lack of error bars on reported metrics and of the quantification of statistical significance more generally. This is the single simplest, most effective, and most needed reform that an editor can insist upon and that a reviewer should look for. Multiple papers here suggest approaches that should be applied [1, 5, 7, 10]. There can be no excuse for a paper on a modeling method to be published claiming one method is superior to another without proper statistical validation. Finally, we hold to the aforementioned second

school of thought i.e. that molecular modeling should be held to the same standards as other fields. As such, our most general recommendation is to report *standard metrics* as a *requirement* and alternates as desired by authors.

Conclusions

Molecular modeling is a relatively young field. As such, its growing pains include the slow development of standards. Our hope for this special issue of JCAMD is that with the help of the arguments made in the contributed papers, the modest recommendations made here will form the kernel of standards that will help us as a community to both improve the methods we develop and to reduce the disparity between reported performance and operational performance.

Acknowledgements The authors gratefully acknowledge the valuable contributions of Marti Head and Terry Stouch, who participated as discussion leaders during symposium on the "Evaluation of Computational Methods" at the 234th American Chemical Society meeting that led to the development of this special issue. Dr. Jain also acknowledges NIH for partial funding of the work (grant GM070481).

References

1. Clark RD, Webster-Clark DJ (2008) J Comput Aided Mol Des. doi:10.1007/s10822-008-9181-z
2. Cleves AE, Jain AN (2007) J Comput Aided Mol Des. doi:10.1007/s10822-007-9150-y
3. Enyedy IJ, Egan WJ (2007) J Comput Aided Mol Des. doi:10.1007/s10822-007-9165-4
4. Good AC, Oprea TI (2007) J Comput Aided Mol Des. doi:10.1007/s10822-007-9167-2
5. Hawkins PCD, Warren GL, Skillman AG, Nicholls A (2007) J Comput Aided Mol Des. doi:10.1007/s10822-007-9166-3
6. Irwin JJ (2008) J Comput Aided Mol Des. doi:10.1007/s10822-008-9189-4
7. Jain AN (2007) J Comput Aided Mol Des. doi:10.1007/s10822-007-9151-x
8. Kirchmair J, Markt P, Distinto S, Wolber G, Langer T (2007) J Comput Aided Mol Des. doi:10.1007/s10822-007-9163-6
9. Liebeschuetz JW (2008) J Comput Aided Mol Des. doi:10.1007/s10822-008-9169-8
10. Nicholls A (2008) J Comput Aided Mol Des. doi:10.1007/s10822-008-9170-2
11. Sheridan RP, McGaughey GB, Cornell WD (2008) J Comput Aided Mol Des. doi:10.1007/s10822-008-9168-9