PERSPECTIVE

Lions and tigers and bears, oh my! Three barriers to progress in computer-aided molecular design

Robert D. Clark · Marvin Waldman

Abstract The computational chemistry and cheminformatics community faces many challenges to advancing the state of the art. We discuss three of those challenges here: accurately estimating the contribution of entropy to ligand binding; reliably estimating the uncertainties in model predictions for new molecules; and being able to effectively curate the ever-expanding literature and commercial databases needed to build new models.

Keywords Curation · Drug design · Entropy · Molecular design · Prediction · QSAR · Predictive uncertainty

Introduction

The road to predicting future scientific and technological developments is paved with spectacular failures and memorable epigrams. Niels Bohr, for example, was fond of saying, "Prediction is very difficult, especially about the future." It is also a corollary of Albert Einstein's quip that, "If we knew what it was we were doing, it would not be called research, would it?" Among the most celebrated examples of erroneous predictions are that flying cars and personal jetpacks would be common-place by now [1, 2]. The 25th anniversary of an event as notable as the launch of the Journal of Computer-Aided Drug Design (JCAMD) compels us to look ahead nonetheless, uncertain though our vision might be. Fortunately, it is a testament to human nature that the direction taken by a scientific discipline is

generally foreshadowed by where the barriers to its progress lie. Here we discuss three of those barriers—understanding the role of entropy in ligand binding, dealing with predictive uncertainty and the difficulty of curating large amounts of data—in hopes that doing so will provide some indication of where the field might be headed in the next 25 years. Our goal here is mainly to raise questions; providing answers to those questions remains for those who come after.

Entropy

When JCAMD began publication, it was easy to believe that understanding ligand–protein interactions was primarily a matter of running sophisticated and long enough molecular dynamics simulations on powerful enough hardware, constrained only by limits on patience and resources. It was no accident that three of the early landmark papers on scoring functions appeared in the journal in the mid-1990's [3–5]. The goal of predicting affinity from ligand and protein structure on a purely mechanistic basis seems less attainable today, though. Empirical methods are often competitive with or superior to mechanistic ones when it comes to predicting affinity within chemical series. In some cases and in the hands of some practitioners, empirical methods are superior, as are some empirical ligand-based methods that ignore the structure of the target protein altogether and are completely empirical in nature. Indeed, the inventor of comparative molecular field analysis (CoMFA)—arguably the archetypal mechanistic modeling approach for ligands—now advocates topomeric alignments that abandon any pretense of relying on realistic ligand conformations [6].

R. D. Clark (✉) · M. Waldman
Simulations Plus, Inc., 42505 10th Street West,
Lancaster, CA 93534, USA
e-mail: bob@simulations-plus.com

There is merit in the argument that the performance of mechanistic models is limited by the inherent difficulty of docking a flexible ligand accurately into a flexible protein target [7], but supplanting a rigid lock and key model of ligand–protein interaction with one involving induced fit begs a fundamental question: if the protein is in a lower-energy state before the ligand inserts itself into the active site than it is after, then where does the extra energy needed to "turn the key" and change the shape of the lock after binding come from? For the process to be spontaneous, there must be a net decrease in free energy (ΔG) above and beyond that gained by the initial insertion of the ligand into the unoccupied binding site.

Somewhat counterintuitively, a major positive contributor to the entropy of binding (ΔS) is *increased* protein fluidity after a ligand binds—in particular, an increase in anharmonic, coupled motions within the complex [8–10]. The coupling is presumably mediated at least in part by polar interactions between residues, which may contribute to the apparent ubiquity of enthalpy-entropy compensation in ligand binding. The resulting entanglement of enthalpy and entropy is, in turn, consistent with recent analysis of isothermal calorimetric data for ligand–protein complexes for which high-resolution complexes are available, work which underscores just how tightly ΔS is connected to the enthalpy of binding (ΔH): Tang and Marshall [11] showed that the residual error in models of the total free energy of formation (ΔG) for such complexes is lower than the residual error in models of its ΔH and $-T\Delta S$ components (T is the absolute temperature).

The scope of the challenge of fully accounting for the affinity of a protein for a ligand based on a static structure of their complex is underscored by the huge difference in stability observed between complexes of time-dependent and time-independent inhibitors with prostaglandin H2 synthase (COX-1). Time-independent binding is fast and readily reversible, whereas time-dependent binding involves a subsequent slow step (minutes) and is nearly irreversible. Nonetheless, X-ray analysis of the complexes shows that the ligands adopt similar conformations and the enzyme adopts "identical" conformations [12]. It would seem that some sort of dynamic effect must account for the differences between the two, and the sort of anharmonic fluidity cited above is a prime candidate. Accounting for such dynamic contributions to ΔS is complicated, however, by their fractal nature (at least in the case of myoglobin, lysozyme and bovine pancreatic trypsin inhibitor) [13], which suggests that they are chaotic or emergent properties of complex systems [14]. If that is true in general, the dynamic contributions will be difficult to deal with using existing mechanistic methodology. Appreciating the chaotic nature of the interaction could, however, help molecular designers understand what makes slow, tight-binding ("time dependent") inhibitors special.

Uncertainty

Regression models of all types have historically been characterized in terms of how well they fit the data upon which they are based, usually by the root mean square error (RMSE), where "error" corresponds to the difference between the observed and predicted response. Prospectively estimating how much the predicted response for a new point in the model's descriptor space is likely to deviate from the observed response—the predictive uncertainty—has proven far more difficult. If a model is based on linear regression and the descriptors are statistically independent variables, classical statistical theory provides tools for calculating how large the deviation from prediction is expected to be for a new observation. Unfortunately, those tools are inadequate when the model is non-linear or the descriptors are correlated, and one of those conditions almost always holds when drug molecules and biological responses are involved.

The desire for a quantitative estimate of predictive uncertainty is not diminished simply because statistical theory fails to provide rigorous tools for providing one, however. Until fairly recently, the only way this need was addressed in quantitative structure–activity relationship (QSAR) analysis was a pragmatic one in which a test set was held back from the model building process and the root mean square error of prediction (RMSEP) was calculated across the compounds in the test set. The QSAR community has recently come to realize that this approach is inadequate when the compound for which a prediction is desired differs too greatly from the compounds in the training set used to build a model or the test set used to evaluate its predictive reliability. That realization has led to the definition of "applicability domains," ranges of descriptors (or, in a few cases, combinations of descriptors) outside of which a model cannot be relied upon to accurately predict responses or properties. The need for a specified and well-defined applicability domain has even moved into the regulatory realm, at least in Europe [15].

Unfortunately, the need to establish explicit measures of predictive reliability has yet to penetrate mechanistic approaches to activity prediction, i.e., "structure-based" molecular design. Perhaps it goes without saying that the absence of appropriate force field parameters for a silicon atom would compromise the predictions produced by a mechanistically based docking program. There also seems to be an implicit assumption in the field, however, that exactly how a given set of carbon, hydrogen, nitrogen, oxygen, sulfur, phosphorous and halogen atoms are assembled to form a new compound is at most incidental to predictive success, and that aggregate measures of retrospective performance on test sets are all that is needed. At the extremes this is clearly an absurd assumption,

especially for methods that have any empirical component at all. It is unlikely that a program trained solely on proteases will accurately dock a steroid into the estrogen receptor, for example, or that a program that has never "seen" a sulfonamide will handle one correctly.

One response to these concerns is that the current pool of known structures for ligand–protein complexes is now so large that it represents a broad enough (albeit markedly biased) sampling of biochemically relevant space that addition of a few new chemical classes or new targets is unlikely to affect the overall performance significantly. Even were this so, it overlooks the point that the *value* of a prediction is a direct function of the novelty of the structure for which the prediction is being made, regardless of whether the "value" in question is social or commercial. Hence, the predictive uncertainty at the edge of the training and test set space is more relevant than the average uncertainty across the entire retrospective structural space. This is one reason that many in industry are turning to rolling tests sets—blocks of compounds assayed since the model was built—to assess the predictive performance of models. The synthetic culture and target areas of interest at a company change over time but they do so fairly slowly, so the compounds made in the last 3 months tend to be different than those made a year ago; not enormously so, but enough that some will lie at the edge of the historical applicability domain. This approach also helps address the point that there may be some *combinations* of descriptors that are relevant for new chemistry but which were inadequately explored in the training and test sets.

The ultimate goal should be making reliable predictive uncertainty estimates for *individual* new compounds. Some exploratory work has already been done in this area utilizing sampling and the variance of ensemble predictions [16, 17], but more is sorely needed if modeling is going to remain relevant to drug discovery. Most of what has been done to date assumes (explicitly or implicitly) that the more similar a candidate molecule is to a molecule from the training set, the smaller the uncertainty in the prediction will be, or that the uncertainty in the prediction is directly related to the prediction errors of training (or test set) molecules that are "similar" to the candidate molecule. The exact meaning of "molecular similarity" can be very sensitive to context however. Molecular similarity, like beauty, is in the eye of the beholder, and the key beholder in this case is a target receptor, not a medicinal chemist and not a computer model.

All similarity-based approaches are prone to failure when presented with what might aptly be called a "phantom variable." Consider, for example, the case of a protein like the PepT1 transporter, which has a strong preference for binding di- and tri-peptides composed entirely of L amino acids ("L-peptides") over peptides that include one or more D amino acid residues; enantiomeric peptides composed entirely of D amino acids ("D-peptides") bind even less well [18]. Hence a 2D QSAR model built using data only for L-peptides is unlikely to correctly predict the behavior of D-peptides. The model will "see" a D-peptide as being within its applicability domain if the enantiomer is, however, and will predict it with confidence just as high (or low) as for the corresponding L-peptide. The same holds true if similarity is based on "independent" descriptors not used to build the model. Fingerprint-based measures of similarity, for example, might place a new D-peptide within the applicability domain of the model, but the transporter is likely to see the situation differently. Nor is the problem limited to 2D QSAR; it extends to 3D methods as well. There will almost always be *something* that distinguishes a given new candidate molecule from the molecules in the training set, and the target may turn out to be sensitive to that difference. Just this sort of situation doubtless accounts for some of the activity cliffs that have received so much attention in the recent literature [19–21]. The problem is further exacerbated for applicability domains by the need to use many descriptors to model target properties for very large data sets. The more descriptors are used, the greater is the likelihood that a candidate molecule will fall outside the limits of one or more of these descriptors—the "curse of dimensionality." This is the reverse of the (reasonable) expectation that the larger the training set, the larger a model's applicability domain should be! It would seem that some measure of molecular similarity that is sensitive to context is needed, as well as recognition that the descriptors needed to construct a model may not always provide enough information to reliably estimate that model's predictive uncertainty for all new candidates.

Curating oceans of data

The rapid increase in computing power since 1986, coupled with the ease of access provided by the World Wide Web and digitization of data sources, has made it easier than ever before to quickly obtain information on an enormous range of topics. Indeed, the authors drew heavily on Internet resources to put together this paper. That data is spread thin, however, in the sense that any single *primary* source generally contains only a few items relevant to any particular topic. The task of pulling dispersed data together once fell to people who wrote review articles. In many areas, that function has been supplanted by databases generated using computer programs. Unlike the authors of review articles—unlike the conscientious authors, at any rate—databases generated by unsupervised search programs tend to agglomerate data rather indiscriminately.

Much of the data contained in any single database is, by its nature, derivative. In many cases, it has been copied over several times since its original abstraction from some primary source. This makes the potential for rapid and broad dissemination of errors enormous. The lack of discrimination involved in automated database construction may be a virtue when it preserves data that might not fit neatly into a human reviewer's preconceptions or might evade their notice, but it becomes a vice when erroneous data fails to be recognized and labeled as such.

The structure of gallamine triethiodide is a good illustrative example where many major databases ended up containing the same mistaken datum. Until mid-2011, anyone relying on an internet search would have erroneously concluded that gallamine triethiodide is a tribasic amine. The error resulted from mis-parsing the common name at some point as meaning that the compound is a salt of gallamine and "ethiodidic acid," identifying gallamine as the active component and retrieving the relevant structure. In fact, gallamine triethiodide is what you get when you *react* gallamine with three equivalents of ethyl iodide (Fig. 1).

It is easy for a name-to-structure conversion program to make this mistake with a common name rather than "the" systematic one, but the "triethiodide" usage is systematic in its own context. It is also arguably quite reasonable—provided one already knows what gallamine is. The same is true for stearates, acetates, palmitates, etc., where esters often get mixed up with salts. When correctly used, such names are correct and unambiguous in context: 4-androstendiol has no basic centers to form a salt with acetic acid, so "4-androstendiol diacetate" has to be an ester. The problem is that the usage does not conform to the *universal* standard expected by software that ignores chemical context.

(As it happens, a simple Web search fails to make it clear what the "correct" IUPAC name for gallamine triethiodide actually is. Several sites, including Wikipedia, [22] give it as "2,2′,2″-[benzene-1,2,3-triyltris(oxy)]tris(N,N,N-triethyle-thanaminium) triiodide," whereas WolframAlpha gives it as "2-[2,6-bis(2-triethylammonioethoxy)phenoxy]ethyl-tri-ethyl-azanium triiodide"[23]. The simpler "1,2,3-tris-(2-triethylammonioethoxy)benzene tri-iodide" is more

informative and fully consistent with the example provided as part of IUPAC Rule C-816.3 [24] but is unlikely to be produced by any automated name generation program that starts by searching for a base name using a prioritized look-up table.)

The authors have been particularly sensitized to this issue by time recently spent curating a data set of cytochrome P450 (CYP) reactions, where typographical errors and misinterpretation of semi-systematic drug metabolite names were all too common. Ironically, the use of such names in the literature (as opposed to fully systematic IUPAC nomenclature, which is more prone to interpretive and typographical errors) often makes it easier to determine the actual structure of substrates and their metabolites. That said, it is the authors' experience that there is no substitute for explicit structural depiction in facilitating validation, something which should be strongly encouraged in all publications, especially when new compounds are introduced.

It is tempting to automate curation itself by accepting the version of any given factoid that appears most frequently on the Web as the correct one. The ease of implementation seems to make the use of such methods inevitable, but the potential for (false) positive reinforcement in such a system may do more harm than good. Indeed, the origin of the quote attributed to Niels Bohr at the start of this article has been addressed in just this way—a process which readily yields the wrong answer [25–27].

Mis-attribution of a quote is arguably inconsequential, but misassignment of structures—either base structures (as for gallamine triethiodide) or tautomeric and protonation states—can severely compromise a model or simulation. This is especially so when the error is systematic, as in the case of steroid esters or ethiodide salts. The extensive use of automated systems for populating public and private databases with standardized structures can severely exacerbate the problem. Misprotonation examples include the misrepresentation of piperazines as diprotonated rather than as monoprotonated species and the conversion of amidinium ions into diprotonated aminals (Fig. 2).

Such unlikely proximal dications are generated when proximity effects—context—are ignored, and docking such species into a binding site that bears a net negative charge will wildly exaggerate the affinity of the compound in most cases.

Typographical and incidental analytical errors creep into databases as well. So long as science progresses at least in part by making constructive mistakes, there will always be a considerable amount of bad data mixed in with the good. Curating such large databases effectively is, however, extraordinarily difficult—and only rarely appreciated when it does get done. The problem is exacerbated by three facts of grant application life in 2011: that support is relatively easy to obtain for compiling new cross-linked databases;
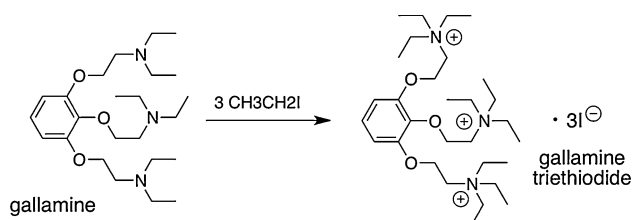
Fig. 1 Conversion of gallamine to gallamine triethiodide

Fig. 2 Common misrepresentations of piperazines (*left*) and amidinium cations (*right*)
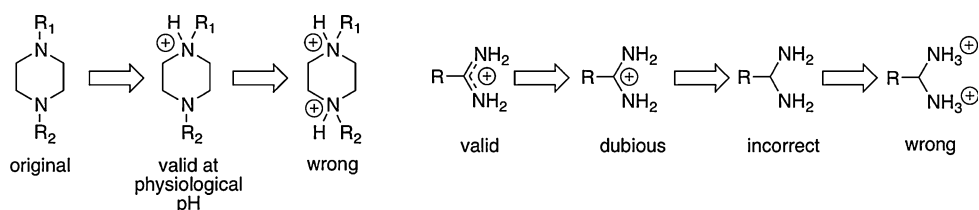
that the perceived value of such databases is based more on the amount of data they include than on the reliability of that data; and that it is very difficult to obtain funding for maintenance or manually supervised curation of existing databases.

This is a remarkable state of affairs for a discipline still recovering from the damage inflicted by the notion that high-throughput screening and combinatorial chemistry programs would produce a groundswell of new drugs by sheer force of numbers. That experience showed clearly that the quality of data is more critical than how much of it there is and underscored the sharp differences in value between data, information and knowledge. One hopes that new compilations will soon be expected to provide an indication of where each datum came from, if not the ultimate primary source. The CYP curation effort alluded to above was greatly facilitated by the data sets being well documented in terms of citation of primary sources. It is also important to cite the actual source from which data was obtained, however. One often finds examples in the literature (and even more often on web pages) where a primary source C is cited, when, in fact, a secondary source B was used which cited source C. Failing to acknowledge such use of indirect data sources readily propagates errors, e.g., when source B incorrectly copies or misinterprets the information in source C.

Conclusion

When Garland Marshall, Andy Vinter and Hans-Dieter Höltje launched this Journal 25 years ago, it was believed by some in the community that computers would ultimately—perhaps sooner rather than later—displace human beings from the drug discovery process entirely. Given that fact, they might well have named it the "Journal of *Computational* Molecular Design." They opted instead for the "Journal of *Computer-Aided* Molecular Design," which proved a prescient choice. It is easier now to appreciate the inherent limitations of computer programs as well as their potential for augmenting and extending the reach of molecular designers. The three hurdles highlighted here—understanding entropy, estimating uncertainty and effective curation of large databases—are daunting but not technically insurmountable. But the psychological and institutional challenges they represent are perhaps equally

forbidding: people naturally tend to focus on the direct and familiar (enthalpy) rather than on the subtle (entropy); they resist quantifying how much they do not know (uncertainty); and they tend to favor increasing quantity at the expense of quality (curation). Those aspects of the challenges, too, can be overcome. If the community does so, the fruit of our labors will be evident to our successors 25 years from now when they read the Journal of Computer-Aided Molecular Design as their cars cruise through the air on autopilot. Or perhaps they will prefer to hear it spoken directly into their minds through their brain chip implant as they try out the latest jet-pack.

References

1. Benford G, the Editors of Popular Mechanics (2010) The wonderful future that never was: flying cars, mail delivery by parachute, and other predictions from the past. Hearst, New York
2. Wilson DH (2007) Where's my jetpack? A guide to the amazing science fiction future that never arrived. Bloomsbury Publishing, New York
3. Böhm H-J (1994) The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. J Comput Aided Mol Des 8:243–256
4. Jain AN (1996) Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. J Comput Aided Mol Des 10:427–440
5. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP (1997) 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 11:425–445
6. Cramer RD (2010) Tautomers and topomers: challenging the uncertainties of direct physicochemical modeling. J Comput Aided Mol Des 24:617–620
7. Cozzini P, Kellogg GE, Spyrakis F, Abraham DJ, Costantino G, Emerson A, Fanelli F, Gohlke H, Kuhn LA, Garrett M, Morris GM, Orozco M, Pertinhez TA, Rizzi M, Sotriffer CA (2008) Target flexibility: an emerging consideration in drug discovery and design. J Med Chem 51:6237–6255
8. Amadei A, Linssen ABM, Berendsen HJC (1993) Essential dynamics of proteins. PROTEINS Struct Funct Gen 17:412–425
9. Zidek L, Novotny MV, Stone MJ (1999) Increased protein backbone conformational entropy upon hydrophobic ligand binding. Nature Struct Biol 6:1118–1121

10. Balog E, Perahia D, Smith JC, Merzel F (2011) Vibrational softening of a protein on ligand binding. J Phys Chem B 115:6811–6817

11. Tang YT, Marshall GR (2010) PHOENIX: a scoring function derived using high-resolution crystal structures and calorimetric measurements. J Chem Inf Model 51:214–228

12. Selinsky BS, Gupta K, Sharkey CT, Loll PJ (2001) Structural analysis of NSAID binding by prostaglandin H2 synthase: time-dependent and time-independent inhibitors elicit identical enzyme conformations. Biochemistry 40:5172–5180

13. Cusack S, Doster W (1990) Temperature dependence of the low frequency dynamics of myoglobin. Biophys J 58:243–251

14. Wikipedia (2011) Complex system (http://en.wikipedia.org/wiki/Complex_system). Accessed 10 Nov 2011

15. Cronin MTD, Jaworska JS, Walker JD, Comber MHI, Watts CD, Worth AP (2003) Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. Environ Health Perspect 111:1391–1401

16. Beck B, Breindl A, Clark T (2000) QM/NN QSPR models with error estimation: vapor pressure and logP. J Chem Inf Model 40:1046–1051

17. Clark RD (2009) DPRESS: localizing estimates of predictive uncertainty. J Cheminf 1:11

18. Biegel A, Gebauer S, Hartrodt B, Brandsch M, Klaus Neubert K, Thondorf I (2005) Three-dimensional quantitative structure-activity relationship analyses of β-lactam antibiotics and tripeptides as substrates of the mammalian H+/peptide cotransporter PEPT1. J Med Chem 48:4410–4419

19. Guha R, Van Drie JH (2008) Structure–activity landscape index: identifying and quantifying activity cliffs. J Chem Inf Model 48:646–658

20. Bajorath J, Peltason L, Wawer M, Guha R, Lajiness MS, Van Drie JH (2009) Navigating structure–activity landscapes. Drug Disco Today 14:698–705

21. Dimova D, Wawer M, Wassermann AM, Bajorath J (2011) Design of multitarget activity landscapes that capture hierarchical activity cliff distributions. J Chem Inf Model 51:258–266

22. Wikipedia (2011) Gallamine triethiodide (http://en.wikipedia.org/wiki/Gallamine_triethiodide). Accessed 2 Nov 2011

23. WolframAlpha (2011) Gallamine triethiodide (http://www.wolframalpha.com/entities/chemicals/gallamine_triethiodide/r2/c6/bn/). Accessed 2 Nov 2011

24. Advanced Chemistry Development, Inc. (2011) Rule C-816 Ammonium compounds (Groups Containing One Nitrogen Atom) (http://www.acdlabs/iupac/nomenclature/79/r79_535.htm). Accessed 2 Nov 2011

25. Guy M (2011) Bohr leads Berra, but Yogi closing the gap. Letter from here (http://letterfromhere.blogspot.com/2006/12/bohr-leads-berra-but-yogi-closing-gap.html). Accessed 30 Oct 2011

26. Denenberg L (2011) Who first said "It is difficult to make predictions, especially about the future" (or one of its many variants)? (http://www.larry.denenberg.com/predictions.html). Accessed 30 Oct 2011

27. ChaosBook (2011) Correct attribution is hard, especially for the past (http://chaosbook.blogspot.com/2010/06/lundskovdk-citater.html). Accessed 30 Oct 2011