J Comput Aided Mol Des (2012) 26:125–134
DOI 10.1007/s10822-012-9541-6

PERSPECTIVE

The errors of our ways: taking account of error in computer-aided drug design to build confidence intervals for our next 25 years

Terry Richard Stouch

Received: 29 December 2011 / Accepted: 4 January 2012 / Published online: 14 January 2012

Abstract The future of the advancement as well as the reputation of computer-aided drug design will be guided by a more thorough understanding of the domain of applicability of our methods and the errors and confidence intervals of their results. The implications of error in current force fields applied to drug design are given are given as an example. Even as our science advances and our hardware become increasingly more capable, our software will be perhaps the most important aspect in this realization. Some recommendations for the future are provided. Education of users is essential for proper use and interpretation of computational results in the future.

Keywords Error · Precision · Force fields · Computational chemistry · Drug discovery · Drug design · Computer-aided drug design · Molecular modeling · Molecular dynamics · Crystal structure prediction

Introduction

It's been 25 years since JCAMD's founding in 1987 and we've published many papers describing QSAR, conformational analysis, drug design, homology modeling, docking, and other technologies. Many different programs have been applied to many different applications and to an ever increasing stream of biomolecular and pharmaceutical problems. There have been some successes and many claims.

Since that time we've seen a dramatic increase in computer power and in new, more user-friendly software. The software is in fact so user-friendly that it seems to make a simple job of some complex science and makes it approachable to novices as well as experts. Computation has become a standard addition to much, perhaps most, of pharmaceutical and biomolecular research.

However, as I write this I'm torn between being impressed with what we can do and have done and disappointed that we can not do and have not done more. Having been in this field for 30 years I must admit that although we've made some inroads I feel compelled to ask if our science has advanced in proportion to our computer power and how much closer we really are to our goals. In fact, I'm disappointed that often it's not even clear that we know where we stand.

Also, I see a dichotomy in the estimation of the value of computational chemistry. It is widely used. Some users are very ready to assume implicitly that the results of the calculations are correct. Some even treat the results with a confidence that goes beyond that for experiment (of course, this is justified in some cases with high level QM—but those are not the applications to which I refer here). Yet, among many experimentalists and non-computational chemists there is substantial skepticism, sometimes almost outright hostility. In fact, often our field as a whole seems to be disparaged by results that don't live up to sometimes unrealistic expectations. Although computational chemistry is often demanded, among many non-computational chemists it is often treated mainly as a curiosity. And, insultingly, many seem not to be surprised when the methods don't work—some even seem to expect it. There is even substantial (and I would say

T. R. Stouch (✉)
Senior Editor-in-Chief, Journal of Computer-Aided Molecular Design
Science For Solutions, LLC, 6211 Kaitlyn Ct,
West Windsor, NJ 08550, USA
e-mail: tstouch@gmail.com

informed and justified) caution among expert computational chemists.

In fact, the skepticism is well-founded on the inconsistencies of our successes—we are wrong perhaps as often as we are right. Could one reason for this be that we seldom have any estimate of confidence in the results of our calculations? Are the domains of applicability of our methods well established? Yet numerical values are frequently given to many significant figures and many users tend to interpret the results literally.

In fact, our confidence in the results is such that many experienced practitioners agree that the true value of computer-assisted drug design (CADD) often lays in the skill of the practitioner as much as in the software. For example, a blinded, controlled test has shown that the prediction of the accuracy of docking studies is best determined by expert humans who are more accurate than present day scoring functions [1, 2]. In addition to running programs, knowledge of what works and what does not and provision of expertise seem to be what truly provides value.

It seems that often the application of computational chemistry has leapt ahead of the underlying science. Sometimes this has been driven by "I wonder if…" and "Let's try…." hypotheses. That is a good pioneering attitude; however, not appropriate for the application of technology. This attitude was endemic in the application of structure-activity relationship studies. Early-on the field saw encouraging success in the derivation of relationships between biophysical properties and carefully determined physical properties within congeneric series of molecules that differed only by small structural changes. However, that success quickly gave rise to much less careful studies that were applied to structurally diverse sets of unrelated molecules and endpoints that were often more semantic than physical (take for instance mutagenicity or toxicity which are broad terms applied to endpoints resulting from often very different chemical and biological mechanisms). Lack of understanding of basic statistical principles led to widely misinterpreted results. The unfortunate consequence was that QSAR's reputation was sullied.

There is an important role and a strong demand for computation to augment, explain, and even go beyond experiment. But how can we increase our true success rates and improve our credibility? How can we make the results of our calculations more reliable? It is likely that our inconsistencies/failures are in part because we do not properly understand the domain of applicability of our methods and the error estimates of our calculations. Understanding these things won't change our success rate, but it will reduce our apparent failures. It will also make clear where we need to improve as well as where we can be successful.

Error and misunderstanding can arise from many sources and the results provided by computational chemistry software arise from a chain of dependencies where each of the links could result in error.

Science → theory → algorithm → software → implementation → adjustable parameters → use → user → interpretation

The science, the true physics, must be expressed as a theory which must be embodied in an algorithm which must then be faithfully translated into software. The software must be implemented into the hardware and system effectively. In most cases there are parameters that must or can be set and can be adjusted. Also, in most cases the user has control of the execution of the software and the end-user (perhaps not even the person who ran the program) interprets the results. Space and time are too short for a proper evaluation of this chain, which must wait for another paper. Later I'll discuss some specifics regarding the first few links.

As I look back at the last 25 years and forward into the next 25 years, it's clear to me that our field is now mature enough that it's time to clean up the dangling ends and determine where we stand, what we can't do and what we can do, and just how well we can do it. Once we know this we can provide confidence intervals to our results and determine what we need to do to improve our situation or if it can't be improved how we can work around it to achieve our goals.

Testing and evaluations of our methods is not a new concept and several organized evaluations have made a good start and continuing effort in helping us understand the quality of our results [1–10]. However, although the results of such evaluations are treated with interest by some, it is not always clear that the lessons are incorporated into routine work. Also one wonders if the results are fully assimilated by non-experts users of software.

Computational chemistry and computer-aided molecular design effectively covers a number of fields of study (including informatics, quantum chemistry, statistics, physical chemistry, bio-physics) and many topics could be used as an example. Here I will concentrate on energies calculated from conventional molecular mechanics force fields (FF) because they underlie so much of our work, many users would like to interpret calculated drug/protein interaction energies, and their fundamentals are generally out of site of expert and novice alike, deep within most software. Like the bridges that many of us cross during our daily commute, we take force fields for granted and assume that they will get us safely where we need to go.

Continual substantial effort by several groups has led to continued improvements and advancements in FF and in this Perspective I do in no way want to imply criticism. These researchers are my heroes; they spend many unsung

hours sweating the details and providing the highest possible quality results. Having participated in this effort myself, I understand the effort that this requires. Still, the aim of a force field is to take a very difficult problem and render it in a fairly simple form and make it fast to compute. As with any instrument or experiment the calculated values do not, and perhaps cannot, perfectly duplicate nature and come with their own inherent error.

Error in calculated drug/protein interactions

The energetics of drug discovery

First it is important to understand the amount of error that our work can tolerate. Since JCAMD deals largely in drug design, I'll address the energetics of drug discovery. Pharmaceutical research generally starts with a 'hit' compound that now-a-days usually at the worst has discernable activity at micro Molar (μM) concentrations. A hit is refined to a 'lead' which is optimized for efficacy and other properties. An activity of nano Molar (nM) is usually sufficient to progress a compound as a drug candidate. That provides a range of 3–4 factors of 10. It's not hard to calculate that one factor of 10 equates to around 1.4 kcal/mol and so 4 factors of 10 constitutes a difference of less than 6 kcal/mol in binding energy. The entire drug design effort operates in this range of energy. Optimization of a lead in order to refine its properties occurs in a much smaller range of often less than a factor of 10, or 1.4 kcal/mol. Differences in experimental results are often interpreted by drug design teams at factors of 5 or even 2, which amounts to as little as 0.4 kcal/mol. Consequently, a drug design team will act in a realm much below 1 kcal/mol, perhaps less than 0.5 kcal/mol.

The key numbers to remember as important goals for CADD are 6 and 0.5 kcal/mol, the total range of interest and the smallest unit of action.

Force fields and approximations

Although their parameterization can be long and involved, conventional force fields are, compared to quantum mechanics, relatively simple things. Still the description of molecular structure and drug/protein interaction is complex since it is composed of the contribution of a great many force field terms and a great many individual (often correlated) interactions. The genesis of FF forms has a long history of which many users are unfortunately unaware. Underpinning these force fields are unavoidable approximations that, although they allow energies and forces to be calculated quickly, (and in my estimation do admirably in many instances) are, none-the-less, still approximations. These include how 1,4 interactions (torsions, dihedrals) are handled and even more importantly calculation of the van der Waals and electrostatic interactions between atoms. For example, the '12' repulsive term of the well-known Lennard-Jones '6–12' equation to calculate the van der Waals interaction energy has no physical basis and was put in place as a computational convenience despite the fact that it is known to be too severe. The sphericity of atoms is an approximation that provides a tremendous calculational convenience that does not hold for tight contacts. Atom centered 'point charges' so convenient, simple, curiously intuitive despite their fallacy, heavily studied, and beloved, were initially put in place as a quick work-around that made results palatable. The particular values for these charges are for the most derived by a statistical fit to molecular electrostatic potential and have several dependencies and their own error of fit. The specific parameters for each atom that are used to calculate atom–atom van der Waals interactions have been based on fitting procedures as well and in some cases, empiricism.

More rigorous approaches are known for many of the approximations; however they increase the complexity of the software, the time and effort of parameterization, and the time required to make the calculations. Also despite their crudeness these approximations work well in many cases and were valuable steps forward that allowed us to move ahead in a reasonable time. In fact, these approaches have worked so well for many of our purposes that we've come to expect more than is realistic.

Where these force fields seem to provide their greatest value is in the determination and maintenance of molecular structure to a reasonable level of precision. In fact, this would seem to be the first and most fundamental goal of a force field. However, where force fields seem to lapse is in the determination of energies. Why is this? Is it predicable? These questions could be addressed at length. However, this discussion will focus only on the derivation of a rough estimate of what might be expected as the lowest possible error estimate for the most rudimentary aspects of drug/protein interaction energies (enthalpies), the van der Waals and electrostatic terms.

Energetics of drug/protein interactions

It's not controversial to say that the average drug will have a molecular weight of about 400 AMUs and be composed of about 40 atoms. A typical drug binding site of a protein will have on the order of 20 amino acid residues composed of on average about 10 atoms providing a binding site of about 200 atoms.

The optimum van der Waals interaction between two atoms will be around 1 kcal/mol perhaps a little less. This value increases rapidly and dramatically as the two atoms

get closer and decreases asymptotically as the atoms get further apart. A reasonable estimate is that each of the 40 drug atoms has 3 close contacts with protein atoms, providing a sum total energy of 1 kcal/mol/interaction × 120 interactions or 120 kcal/mol. A reasonable estimation of more distant, second shell, interactions might be 5 interactions per drug atom with a modest energy of 0.25 kcal/mol providing $40 \times 5 \times 0.25 = 50$ kcal/mol. Consequently, the total van der Waals energy of just the closest contacts will total 170 kcal/mol.

Errors in calculated interaction energies

What is the error in this value? Calculation of the error in van der Waals interaction is the most direct to make due to the established lower limit in the accuracy of the combining rules [11, 12]. van der Waals interactions are typically computed using the Lennard-Jones or similar equation employing parameters that have been derived for each atom type and whose joint interaction is determined using a 'combining rule'. Early such rules were reported by Shneior Lifson as having errors of as much as 60% when applied to the experimental data for the interaction of atoms of the rare gases. The best possible error estimate for more recent and more complex formulations is 1% [11].

Provided that current CADD software uses these most recent and rigorous combining rules and assuming this 1% was accurate, then for our example the error due to the most precise combining rule alone would be 0.23 kcal/mol. Remember that this is an estimate only for the first 2 shells of the drug/protein interaction and that only for the optimum interaction energies. No higher energies due to close contacts have been considered (and recall that the repulsive term is not well-determined). The non-convergent nature of the dispersion forces have not been summed which although of progressively lesser energy never reach zero even as the number of interactions increases with distance. Also, it is important to keep in mind that this error estimate is only for the error in the combining rules and does not include any error in the atomic parameters themselves, which for drug/protein interactions are not as well determined as those for the rare gases for which the 1% error was determined.

In fact, informed sources suggest and agree that when all these factors are taken into account the error in calculated van der Waals interactions are easily closer to 10%, which yields a numerical value of about 1.5 kcal/mol. This value alone swamps the 0.5 kcal/mol precision required for optimization and challenges that for the entire range of drug discovery of 6 kcal/mol. Although above I address only the enthalpy, it is interesting to note that even small differences in the Lennard-Jones terms have been shown to have significant influence on hydration *free energies* of small molecules [13].

It is sobering to think that van der Waals energies are small in comparison to electrostatic interactions. Note that given a dielectric of 1.0, two atoms with a conservative 0.2 electron unit charge each at contact distance of 3.0 A will have an energy of interaction of 4.4 kcal/mol (applying Coulomb's law, $332 \times 0.2 \times 0.2/(3.0 \times 1.0)$). Should the error in this term only be the same1% (which I suspect is dramatically low) that we applied to the combining rules, then the error would be 0.04 kcal/mol. The error would be introduced by the approximation in the statistical fit of the derived charges and also by the inaccuracy of the dielectric of the medium (provided that the commonly-used assumption of dielectric continuum is appropriate at this microscopic level of detail; it is commonly used that way). Summation over only the estimated 120 close contacts would yield a total possible error of 0.44 kcal/mol. I won't take this calculation to further shells but note that whereas van der Waal interactions attenuate rapidly (distance to the 6th power) electrostatic interactions attenuate linearly with distance and, like the dispersion term, are non-convergent. Treatment of the dielectric could be a substantial source of error. Two ions of unitary charge Å apart in a medium of dielectric of 1.0 will have a 33.2 kcal/mol interaction energy. But at a dielectric of 80 that value would be only about 4 kcal/mol. How precisely do we know the dielectrics of our systems? What is the correct value for any particular calculation? Which value between 4 and 33 is correct?

It is important to be aware that the point charges are approximations of the electrostatic interactions between atoms. The rigor of point charges has been questioned on a practical level by one of their pioneers, Donald Williams [14]:

> Traditional net atomic charge models were found unsatisfactory for representing the molecular electric potential (MEP) of n-alkanes ··· Fitting the MEP with potential-derived net atomic charges (PD charges) gave errors ranging from 51 to 62% with the same basis set. The use of larger basis sets, inclusion of electron correlation, use of more MEP data points, or relaxation to optimal structural geometry did not improve significantly the representation of the MEP by net atomic charges. ··· To improve the representation of the MEP of n-alkanes, augmentation of the model with nonatomic sites … additional charge sites located between hydrogens, on a line bisecting the CH2 group, achieved fits to the MEP with errors reduced to 8% or less, except for n-butane, where the fitting error was 16%. …

Indeed, more recently it has been reported that what might be considered as small changes in approach to the calculation of partial charges or in small perturbations of the charges themselves make a significant difference in computed free energies of binding [15] or hydration [16]. Further, "unpublished data … suggests that extremely small changes in partial charges (i.e. at the level of 0.02–0.05 e) can in some cases lead to changes in hydration free energies at the 1–2 kcal/mol level" [17].

At this point, it is even difficult to estimate the error in calculated electrostatic energies. However, they surely add quite a bit to our previous estimate of error of 1.5 kcal/mol for van der Waals energy.

The intent of this discussion is simply to provide a low-ball estimate of the minimum expected error in calculated interaction enthalpies. Unmentioned are issues of pairwise additivity, polarizability, and charge transfer, among others. Except for noting some references, I do not address *free* energies of interactions which will include (de) solvation energy, entropy, and the response of the system as a whole. Those properties are other sources of error. However, as we can see just for our attenuated estimates, even the combining rules alone provide a not insignificant error and the estimated 10% error in van der Waals energies as a whole become eye-opening. When potential errors from the electrostatic terms are considered, it's not hard to convince ourselves that the error in a calculated drug/protein interaction enthalpy alone might approach or exceed 6 kcal/mol—the total range of drug discovery.

Intramolecular terms

The focus of the foregoing discussion was intermolecular interaction energy and so did not consider the error in any of the *intra*molecular force field terms. However, they are worth considering for other reasons. We might be forgiven for hoping that bond stretching and angle bending are 'hard' enough, physical enough, and sufficiently determined by experiment that they will not provide a major source of error. However, each drug molecule has on average 6 rotatable bonds. Each amino acid residue has an average of 4. The barriers to rotation can range from nearly non-existent for some single bonds to 20 kcal/mol for the peptide bond. The most common range is between 1 and 3 kcal/mol. Few molecules are at a zero energy conformation and most will have some torsions that climb up the 'well' of the potential. The shapes of some of those wells can be complex and difficult to duplicate with the analytical functions that are usually used. The errors in these terms are difficult to determine, but that should not stop us from considering their impact. What is the likelihood that the calculated energies will be off by even a quarter kcal/mol? What will be the effect?

But don't the errors cancel?

When faced with the prospect of the errors we discuss here, the overused term 'cancellation of errors' is often provided as the panacea. Are we to believe that the error in contributions of hundreds, perhaps thousands, of terms and interactions in a drug/protein complex will somehow just disappear or be reduced to insignificance uniformly and reliably? We need to recall our P-Chem error analysis calculations. When dealing with formal error analysis, one must carry along the error from all sources. The argument of cancellation of errors provides some comfort and perhaps gives some explanation to why our results are not worse. No doubt they must cancel sometimes; likely this is why we *sometimes* see trends for series of congeners. However, even this is not reliable. In the application of force field calculations, we don't know where or how error cancellation will occur. I submit that this is particularly true for complex, asymmetric, anisotropic biological systems. Nature does not throw away kcals or even fractional kcals. She keeps track and sums them all up. So must we. But Nature does not make errors as we do when we try to duplicate her work. We need to track our inaccuracies. Of course if we can somehow convince ourselves that all errors cancel, then we have to find some more mysterious way to explain our failures.

It's amazing-but only to a point: case in point, prediction of small molecule crystal structure

In fact, given these numbers, it's actually amazing how well it all works for determination of structure. I'm certainly no stranger to using force fields and I'm regularly impressed by how well force fields can duplicate the structure and conformation of isolated small molecules in vacuo, having routinely compared a number of different force fields to experimentally and quantum mechanically determined structures. Additionally, I marvel at how well force field driven molecular dynamics (MD) simulation can duplicate and maintain biological systems, including highly fluid lipid bilayers [18, 19]. My assumption is that they work so well partly because once atoms are connected by the strong bond stretching and angle bending forces, there are only so many ways to pack connected balls together in a box.

However, we biomolecular modeling types should not be too sanguine. Such force fields are not universally successful even for structure or other physical properties. For example, the materials community has known for years that achievement of experimental densities of small molecule organic fluids can still be a challenge for many force fields [20]. Densities are completely dependent on the nonbonded parameters and the pressure of a system.

A more detailed case in point is the prediction (and even maintenance under optimization and simulation) of small molecule crystal structure. Many attempts over years to reproduce this high precision data have shown that at this precision even structural data cannot be adequately reproduced with conventional FF. Day et al. [5] note:

> Much of the improved success over previous blind tests can be associated with improvements in how we calculate the relative energies of putative crystal structures. All successful predictions in this blind test were achieved by going beyond standard force fields, although several very different approaches are represented in the successful methods… The evidence here and in the published literature strongly suggests that off-the-shelf force fields will not be generally successful for the final energy ranking in crystal structure prediction.

Although the accuracy and precision of our force fields seem to be good enough to reproduce many aspects of biomolecular structure, the data to which we compare is much lower precision than that of small molecule crystal structures. *At this point, our force fields seem to have reached their limit even for structure.*

In fact, the assumption held so dear to many of us due to its ease of use, that atoms are spherical, has been questioned for decades. Day questioned this for chlorine atoms [22] and Donald Williams, having used nonsphericity for many years in describing aromatic hydrogen atoms, remarked in one of his last papers [21]:

> Molecular packing analysis of crystalline dichlorine (Cl^2) shows that the crystal structure is compatible with an isotropic intermolecular force field that includes polar flattening of the exchange repulsion energy and a five center distributed monopole representation of the molecular electric potential. Polar flattening largely accounts for the short intermolecular contacts in the crystal, and the molecular electric potential is critical in determining the molecular orientation and space group. Neither an atom-centered model for repulsion nor a molecular quadrupole model for the electric potential is adequate to describe the force field of this molecule

The same physical laws and forces that govern small molecule crystal structures and the density of liquids govern biomolecules and drug/protein interactions too. It is important to note that lower precision data of those latter fields does not mean that the same effects seen in small molecule crystal structures are not present in biomolecular structure or that we can ignore them. It just shows that duplication of the less precise biomolecular data is less of a challenge.

Beyond energies

Although I addressed calculated energies above, what are the consequences of force field error for interpretation of MD simulations and molecular conformations? Recent, exciting work in long timescale MD approaches a time span that impacts biological significance. However, many biological motions of proteins appear to be governed not by simple changes of individual torsions but by a large number of concerted, or perhaps coincidental, small changes in multiple rotors and inter-atomic interactions. How then will even small errors in torsional profiles and small errors in atom–atom interactions affect an MD trajectory, the pathway of movement, and the timescale of an event? Are the enticing movements that we see in these simulations truly the waltz of nature or are they the unpracticed movements of a dancer in training?

In terms of generating molecular conformations, many force fields are well able to adequately determine the lowest energy conformations. However, we know that environment and inter-molecular interactions, such as that of drug-protein interactions can result in higher energy conformations. Many of our applications utilize conformational sampling for just this purpose. The determination of the minimum of a conformational well is easier than determination of its shape. (For years even something as fundamental as the barrier to rotation around a carbon–carbon single bond in alkanes was several kcal/mol too high in most force fields.) How then will inaccuracy in torsional profiles affect the distribution of conformations resulting from conformational search?

Errors in force fields, reprise

It is absolutely not surprising that calculated energies or free energies of binding often do not correlate with binding affinities or other measures of activity. It is likely that the error in the force field does not allow interpretation to the level of precision that we need. It is straightforward to recognize that the minimum error in calculated drug/protein interactions is likely comparable to the entire range of pharmaceutical discovery. Consequently, it would be inexplicable if this did work. Often, attempts are made to achieve the perceived experimental values by adding approximations of entropy or solvation—but these approximations cannot make up for error in the calculated energy. Further, how much error will these additional terms bring with them?

There is nothing inherently *wrong* with the force fields. Considering their simplicity (relative to quantum mechanics or the force fields required to duplicate small molecule crystal structures) and the data available for their development, they perform admirably. They owe us nothing. Where

things go awry is in what we seem to expect of them, how we use them, and how we represent the results. To say that computational chemistry has somehow failed because we cannot yet calculate accurate energies is to ignore the underpinnings of the science that we are using.

Perspective and hopes for the future

I've just spent several pages talking about the past, but this Perspective is about the future. I think that we as a field are now at a point where, like an artist, we have painted with a broad brush and now it's time to fill in the details. This is essential not only for our credibility, but also to help us know what we can and cannot do, what we need to fix, what can't be fixed, and consequently, what we need to work around. I hope that what the future will bring is an effort not to be ready to claim success, but rather to determine just where we can be successful and to what degree. We need to try to reduce error when possible, be resigned to the error that is unavoidable, and find work-around when necessary to get us to our goals. Although my examples concern force field calculations, I do not limit my Perspective on the future to that technology only but address all areas of computational chemistry and computer-aided drug design.

First, determine where we stand

Needless to say based on the discussion above, I hope that the very near future will see us routinely determine and deploy the error in our calculated values and employ this essential information in their use. We should interpret the results of our calculations in a way that is appropriate to the science and not force the interpretation to abide by the way we wish things to be.

Consequences

What are the consequences of understanding the error of our calculations? One consequence is that it might mean not giving an answer or that we cannot give an answer to the desired precision and accuracy. Hopefully it will mean that we will be truly "wrong" less frequently and that when we are "wrong" it will be qualified. Someone once insisted that "any answer is better than no answer." I strongly disagree and say that "no answer is better than a wrong answer" or an answer that might mislead. *Primum non nocere* (First, do no harm). However, the best answer is one that is given with an estimation of confidence.

Hopefully, the realization that the desired accuracy or precision is not available will spur more innovation and demand for more fundamental science and enhanced approaches. We've seen that already in some cases, for example in terms of the cooperation of experimentalists to provide data for evaluations, tailored scoring functions, and the recognition that a human expert's eye is an essential adjunct to docking and scoring.

Further, I also hope that this information will help journals develop consistent, firm, editorial policies that will guide authors and help reviewers and editors ensure publication of science that adds increasing value.

Evaluation and standards

The challenges, competitions, and evaluations mentioned above [1–10] have been a good place to start in terms of evaluating the science that can be done. But how do users know if the software they use meets the highest standards? How do users know if their software is working as expected? There are ad hoc tests that many run; some more carefully than others. Should our field—and should users—require results of standard tests be provided and should they, themselves, run the tests? As repellent as it might sound, could the future see some sort of "good computing" certification endowed on software that meets certain community expectations?

Our software

Our problems are complex. The science we apply is complex. Often even trained computational chemists are not completely versed in the science they are computing and often are not intimate with the inner workings of the software that they employ. A lot of our future will rest on our software, more so than our hardware, and even more so than our science if our software is not true to it. I hope that software developers will lead the way and will provide error estimates and some indication of confidence and, if possible, just plain tell us if something can't be done. This could require some substantial additional research, effort, and code; but in the end the results will be more reliable and less prone to inappropriate interpretation. In some cases this should be straightforward, such as error in prediction of regression models. QSAR models have error of fit. Why do so few reported results not at least reflect the error in the training, much less the prediction sets? Why are errors in coefficients routinely ignored? Error in force fields is more difficult to implement, but the technology is available now.

Fortunately, a vanguard of progress is being made in several laboratories to incorporate error, prioritization, and probability in the interpretation of both computational as well as experimental results [9, 24, 25]. Further, tools that help us understand the results, their context, and their proper use are very possible now and occasionally are

already implemented. For example, in my own laboratories and others of my acquaintance, software that estimates physical properties also provides the experimental result, if available, and/or provides the experimental and predicted values of related compounds. pKa predictions are a good example. If an experimental value is known, why not show it? Also, we know that the pKa of atoms in some complex molecules are more accurately predicted then others and this can be derived from the training sets of compounds on which the predictive approach is based. Unfortunately, at least one or two programs present numbers to the same precision regardless of the likely uncertainty and this important distinction is missing. Yet, its provision would improve the value of the software in that the likelihood of misinterpretation would be reduced and consequently it would reduce the likelihood that the software will earn a poor reputation. Also, it would help to determine the most effective use of the result. Some applications might require precision to less than one log unit and in some cases prediction will fall short of this goal. However, analytical labs interested in determining an approximate value in order to determine the required experimental range of measurement are often satisfied with precision of several log units.

Docking software could cluster compounds of like structure and, with the proper interface, could give a list of related compounds in a company's database as well as information of availability of a sample of the compound, synthetic procedures, and physical and possibly biological properties as well. This would provide a means to prioritize hits based on practical information rather than simply piling a long list of molecules on a chemist's desk which leads to biased and uninformed selections.

Future software

But confidence intervals, provision of information, and making things simpler are only the first step. We want to solve problems yet we often try to shoehorn them into particular technologies. Can our software go beyond simply providing numbers? Although there are exceptions, most current software is task oriented rather than goal oriented: an energy is calculated, a molecule is docked, a protein sequence is modeled. Although it is a leap in complexity, could software become *goal* oriented? For example, perhaps a homology model would *not* be built if the model could not be built to the precision required for the ultimate use, say the accurate prediction of a drug/protein free energy of interaction. Maybe a QSAR program would not provide a model if the correlation was poor, the number of descriptors is too large, or the correlation was driven by a few points of high leverage. (Although some users might not like this, it is not apparent that all users have training in even basic statistics.)

But will computation get us where we need to go? Will we ever be able to rapidly calculate accurate numbers to the required accuracy and precision? More importantly, will it get us where we need to go *now*? Perhaps not. If not, could clever algorithms and approaches be designed to help us to meet our goals despite this? Might our future drug design software in addition to our current algorithms to calculate physical properties contain a conglomerate of approaches all assembled toward our goals: expert systems, decision trees, QM where appropriate, database lookup of known information, and table lookup of pre-calculated values? For example, why trust quickly calculated hydrogen bond strengths, pi-pi, and pi-ion interactions when we know that they can require quite high levels of quantum chemistry to determine [23]? Why not pre-calculate them and use table lookup? Might future software make requests for expert intervention, recognize when key data is not available, the need for more data, and even make recommendations for further experiments? Could the software inform a user if a desired model was already available?

I very much hope to see the day when our software routinely calculates precise physical chemical data that will solve our problems from first principles. However, insofar that we need answers now, might we at least temporarily relax our idealism and our passion for the elegance of such approaches and resort to more practical solutions? Perhaps we could employ knowledge engineers to mine the approaches used by those who "have the knack"; for example those who routinely do better than computation in predicting docked poses of drugs. Also I wonder if we try to do too much. Do we have to grab the brass ring to be useful? For example, over the years I've been approached by many companies (some, perhaps most, no longer in business) that have promised to find "the one compound" that will be "the" drug. Some (again, I've never heard from most of them again) have been very proud of progressing a drug discovery program with only a handful of compounds. That is an ambitious task. In fact, I submit that it would be even more useful to recommend a high quality set of several hundred compounds that would span pertinent and synthesizable chemical space. If truly this were done it would realize easily a 10-fold reduction in the number of compounds historically made for a drug discovery program. Further, it would provide reassurance, backing for patents, and backup and second generation compounds. This is just one example, but the intent can be applied to many applications: let's not pursue some ultimate but potentially unobtainable goal and miss the more tractable goals along the way.

Education

Computational chemistry software is often complex scientifically and usually complex algorithmically, can be run

with different conditions and parameters, and produces results that are subject to interpretation. Yet skilled software developers have made many of the programs impressively easy to use. However, even experts can have difficulty knowing how to most effectively run software. Can novices truly understand the results or the implication of default parameters when even experts might be uncertain? Is there a way to provide education so that all users know at least the fundamentals of the science that they employ? Do novices run mass spectrometers or gigaHertz NMRs? Are software developers compromised by the fact that users might not be able to properly use the software or evaluate the results of their programs?

Computational chemistry is a routine adjunct to experimental studies. Many schools offer elective courses in molecular modeling basics. Could this evolve into a series of required classes in the future? Would this include fundamental classes in physical chemistry and statistics? (Would they be trained in error analysis?) Could we even see experienced practitioners required to take routine refresher courses?

Critique of experiment

Let's not ignore the fact that experiment itself comes with error. Yet, often the data that we use to develop and parameterize our methods has not been critiqued, although this is changing. For example, a detailed evaluation has been made of the placement and conformation of ligands in protein/ligand X-ray crystallographic complexes [26]. And for some reason modelers work very hard to fit experimental data as precisely as possible. Contrary to the message I send in the rest of this paper, I think it's fair to suggest that in some cases we are being too hard on ourselves and our methods when we insist on strict comparison to potentially faulty experiments or fail to take the error in the experimental results into account. Many of us have developed QSAR models whose initially apparently erroneous predictions were later validated by additional experiment. After many hours of discussion with experimentalists and examination of large quantities of pharmaceutical data, I can safely say that the actual errors of individual measurements can routinely be much larger than the ideal error of measurement and can often range up to 10-fold for apparently equivalent experiments done between sites even within the same company [27]. However this is a topic for another paper.

Conclusion

I address the perception of the inconsistencies and failures of our field to-date and some would imply that they are endemic to and inseparable from our results. This is not true. I submit that if we've failed to-date, it is in not knowing where we are, what we can and can't do, and the accuracy and precision of our calculations. Some of the disappointments of computational chemistry are due to over-expectations of what our methods should provide. Like poor mule drivers, we have a tendency to push our methods beyond their limits, kick them when they are down, and curse when they don't continue to work as hoped.

We are still a somewhat young and evolving field and any oversight is understandable. However, we are also mature enough to be introspective and honest with ourselves in terms of expected and actual errors and we should incorporate this knowledge into our routine work. This information needs to be made clear to users, expert and novice alike. We should not leave it to interpretation and experiment to provide some binary estimate of our success or failure. If nothing else, that is unfair to our field. *This* I *see as an essential goal of the earliest part of the next 25 years and a continuing effort for the rest of that time and beyond.*

Despite my criticisms, in my opinion the future is bright. Our field has often undertaken problems that even experiment cannot completely embrace. We've had successes and show at least incremental progress, albeit slow in some cases. We have the history and the computing power to understand what we need to do to improve and to do so on a continuing basis. Our methods are highly espoused by increasingly large numbers of users. Our field is here to stay.

We will see improvements on many fronts. The competitions and challenges that are mentioned above and hopefully more to follow are helping to set benchmarks and alert us to problems. Certainly as computer speed increases we will have the ability to run large, increasingly more accurate quantum mechanical systems of increasing complexity and to perform simulations long enough to get statistics that will allow increasingly more realistic comparison to experiment. As we learn our weaknesses we can continue to team with experiment to get more data with which to improve our science. In fact, the results of our computations will highlight the need for more experiments. Also, experimentalists will continue to increasingly rely on computational chemistry. There is lots of room for our software to become smarter and more resourceful and there are ample talented developers and software companies that can make it happen. Hopefully, also, the increased reliance on computation will be accompanied by an increase in the education of users in physical chemistry, statistics, and computation.

Acknowledgments Many thanks to Marvin Waldman, Anthony Nicholls, Robert Clark, and Yvonne Martin for an insightful review of and contributions to the manuscript. Thanks to David Mobley and

John Chodera for sharing their results. The author is indebted to Donald Williams, Marvin Waldman, Sarah Price, Carl Ewig, Arnold Hagler, Shneior Lifson, Peter Kollman, Jay Ponder, Alexander MacKerell, Bernard Brooks, and many others for teaching him about the underbelly of force fields. My appreciation to Anthony Nicholls, Marvin Waldman, William Swope, Julia Rice, Richard Friesner and Thomas Halgren for always invigorating discussions.

References

1. Head MS (2010) Docking: a domesday report. In: Merz KM, Ringe D, Reynolds CH (eds) Structure and ligand-based drug discovery. Cambridge University Press, Cambridge, pp 98–119
2. Skillman A, Geballe M, Nicholls A (2010) SAMPL2 challenge: prediction of solvation energies and tautomer ratios. J Comput Aided Mol Des 24(4):257–258
3. Carlson HA, Dunbar JB (2011) A Call to Arms: what you can do for computational drug discovery. J Chem Inf Model 51(9): 2025–2026
4. Bardwell DA, Adjiman CS, Arnautova YA, Bartashevich E, Boerrigter SXM, Braun DE, Cruz-Cabeza AJ, Day GM, Della Valle RG, Desiraju GR, van Eijck BP, Facelli JC, Ferraro MB, Grillo D, Habgood M, Hofmann DWM, Hofmann F, Jose KVJ, Karamertzanis PG, Kazantsev AV, Kendrick J, Kuleshova LN, Leusen FJJ, Maleev AV, Misquitta AJ, Mohamed S, Needs RJ, Neumann MA, Nikylov D, Orendt AM, Pal R, Pantelides CC, Pickard CJ, Price LS, Price SL, Scheraga HA, van de Streek J, Thakur TS, Tiwari S, Venuti E, Zhitkov IK (2011) Towards crystal structure prediction of complex organic compounds—a report on the fifth blind test. Acta Crystallogr Sect B, 67. doi: 10.1107/S0108768111042868
5. Day GM, Cooper TG, Cruz-Cabeza AJ, Hejczyk KE, Ammon HL, Boerrigter SXM, Tan JS, Della Valle RG, Venuti E, Jose J, Gadre SR, Desiraju GR, Thakur TS, van Eijck BP, Facelli JC, Bazterra VE, Ferraro MB, Hofmann DWM, Neumann MA, Leusen FJJ, Kendrick J, Price SL, Misquitta AJ, Karamertzanis PG, Welch GWA, Scheraga HA, Arnautova YA, Schmidt MU, van de Streek J, Wolf AK, Schweizer B (2009) Significant progress in predicting the crystal structures of small organic molecules—a report on the fourth blind test. Acta Crystallogr Sect B, 65. doi:10.1107/S0108768109004066
6. Fernández-Recio J, Sternberg MJE (2010) The 4th meeting on the critical assessment of predicted interaction (CAPRI) held at the Mare Nostrum, Barcelona. Proteins: struct funct and bioinform 78(15):3065–3066
7. Guthrie JP (2009) A blind challenge for computational solvation free energies: introduction and overview. J Phys Chem B 113(14): 4501–4507
8. Moult J, Fidelis K, Kryshtafovych A, Tramontano A (2011) Critical assessment of methods of protein structure prediction (CASP)—round IX. Proteins: Struct Funct and Bioinform 79(S10):1–5
9. Muchmore SW, Debe DA, Metz JT, Brown SP, Martin YC, Hajduk PJ (2008) Application of belief theory to similarity data fusion for use in analog searching and lead hopping. J Chem Inf Mode 48(5):941–948
10. Warren GL, Andrews CW, Capelli A-M, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2005) A critical assessment of docking programs and scoring functions. J Med Chem 49(20):5912–5931
11. Tang KT, Toennies JP (2003) The van der Waals potentials between all the rare gas atoms from He to Rn. Chemphyschem 118(11):4976–4983
12. Waldman M, Hagler AT (1993) New combining rules for rare gas van der Waals parameters. J Comput Chem 14(9):1077–1084
13. Mobley DL, Bayly CI, Cooper MD, Shirts MR, Dill KA (2009) Small molecule hydration free energies in explicit solvent: an extensive test of fixed-charge atomistic simulations. J Chem Theory Comput 5(2):350–358
14. Williams DE (1994) Failure of net atomic charge models to represent the van der Waals envelope electric potential of n-alkanes. J Comput Chem 15(7):719–732
15. Mobley DL, Dumont E, Chodera JD, Dill KA (2007) Comparison of charge models for fixed-charge force fields: small-molecule hydration free energies in explicit solvent. J Phys Chem B 111(9): 2242–2254
16. Mobley DL, Graves AP, Chodera JD, McReynolds AC, Shoichet BK, Dill KA (2007) Predicting absolute ligand binding free energies to a simple model site. J Mol Biol 371(4):1118–1134
17. Mobley DL (2011) Personal communication on the sensitivity of hydration free energy to charge variation
18. Tejwani RW, Davis ME, Anderson BD, Stouch TR (2011) An atomic and molecular view of the depth dependence of the free energies of solute transfer from water into lipid bilayers. Mol Pharm 8(6):2204–2215
19. Tejwani RW, Davis ME, Anderson BD, Stouch TR (2011) Functional group dependence of solute partitioning to various locations within a DOPC bilayer: a comparison of molecular dynamics simulations with experiment. J Pharm Sci 100(6): 2136–2146
20. Case FH, Chaka A, Moore JD, Mountain RD, Ross RB, Shen VK, Stahlberg EA (2011) The sixth industrial fluid properties simulation challenge. Fluid Phase Equilib 310(1–2):1–3
21. Williams DE, Gao D (1997) Effects of molecular electric potential and anisotropic atomic repulsion in the dichlorine dimer and crystalline chlorine. Inorg Chem 36(5):782–788
22. Day GM, Price SL (2003) A nonempirical anisotropic atom-mâ'atom model potential for chlorobenzene crystals. J Am Chem Soc 125(52):16434–16443
23. Sherrill CD, Sumpter BG, Sinnokrot MO, Marshall MS, Hohenstein EG, Walker RC, Gould IR (2009) Assessment of standard force field models against high-quality ab initio potential curves for prototypes of π–π, CH/π, and SH/π interactions. J Comput Chem 30(14):2187–2193
24. Segall M, Champness E, Leeding C, Lilien R, Mettu R, Stevens B (2011) Applying medicinal chemistry transformations and multiparameter optimization to guide the search for high-quality leads and candidates. J Chem Inf Model 51(11):2967–2976
25. Swann SL, Brown SP, Muchmore SW, Patel H, Merta P, Locklear J, Hajduk PJ (2011) A unified, probabilistic framework for structure- and ligand-based virtual screening. J Med Chem 54(5):1223–1232
26. Nicholls A, McGaughey GB, Sheridan RP, Good AC, Warren G, Mathieu M, Muchmore SW, Brown SP, Grant JA, Haigh JA, Nevins N, Jain AN, Kelley B (2011) Molecular shape and medicinal chemistry: a perspective. J Med Chem 53(10): 3862–3886
27. Stouch, TR (2011) The intricacies of pharmaceutical data: what is required to properly understand it. OpenEye Scientific Software, EuroCUP meeting, Dublin, Ireland, September 2011, http://ww.eyesopen.com/events/eurocup5