J Comput Aided Mol Des (2012) 26:51–56
DOI 10.1007/s10822-011-9514-1

PERSPECTIVE

Computer-aided molecular design under the SWOTlight

Darren V. S. Green · Andrew R. Leach ·
Martha S. Head

Received: 17 November 2011 / Accepted: 2 December 2011 / Published online: 15 December 2011

Within the computational chemistry community at GSK, the authors of this perspective are collectively referred to as "that DAM group", directors of computational chemistry groups in the Computational and Structural Chemistry department. Amongst them, the authors have more than 75 years combined experience in computer-aided drug discovery and more than 30 years combined experience as comp. chem. group leaders. We also belong to that generation of scientists whose careers are approximately the same age as this journal; indeed, one of us published his first scientific paper in the first issue of this JCAMD.

Our reflections in this perspective are based in part on a Strengths, Weaknesses, Opportunities, Threats (SWOT) analysis undertaken by the computational chemistry groups at GSK. Our analysis was largely intended for internal consumption as part of our annual objectives-setting process, but we believe that the application of the SWOT approach to the wider discipline of computer-aided molecular design would also be an instructive exercise. The rest of this perspective will be organized based on a re-ordering of the SWOT themes: Threats, Weaknesses, Strengths, Opportunities. As we reflect back over the past 25 years of our careers as practicing computational chemists, what has been achieved by our scientific community? Has progress been made? How will the discipline change over the next 25 years?

Threats

The environment within which we work is undergoing radical change. Within the 25 years lifetime of this journal, we have seen large changes in the pharmaceutical industry and in the role of computational chemistry within that industry. Once-great companies have disappeared through shrinking pipelines and wholesale mergers and acquisitions, while in that same time period new companies have come into being and new areas of research have opened up. We hear predictions, ideas, fantasies even about how Pharma will change to face the very real challenges of the looming patent cliff and the ballooning costs of discovering new chemical entities. We do not pretend to have any new insights on how to meet these challenges, but we do feel quite confident predicting that even 5 years from now the pharmaceutical industry will look very different than it does today. We live on the cusp of that looming change, and the uncertainty and instability inherent in that metastable state can seem a threat to computer-aided molecular design as a discipline and to each of us as individuals working in that field. That expected change can seem an overwhelming threat, but we would argue that the current situation offers real opportunities for computational chemistry to thrive as a discipline and for computational chemists to capitalize on our relevance, resilience, and impact on drug discovery.

The computational chemistry community is vibrant and is a significant asset. However, we should be aware of long-term threats to our community "memory". Compared to other areas of science, computational chemistry is a

D. V. S. Green · A. R. Leach
GlaxoSmithKline Medicines Research Centre,
Gunnels Wood Road, Stevenage, Hertfordshire, UK
e-mail: Darren.VS.Green@gsk.com

A. R. Leach
e-mail: Andrew.R.Leach@gsk.com

M. S. Head (✉)
GlaxoSmithKline Pharmaceuticals, 1250 South
Collegeville Road , Collegeville, PA 19426, USA
e-mail: Martha.S.Head@gsk.com

relatively young discipline, and the number of scientists involved in computational chemistry is relatively small. This has enabled a strong community to develop involving the entire spectrum of computational chemists, whether they work in academia, industry or in software development. Several conferences have now been established for many years under the auspices of learned societies such as the Molecular Graphics and Modelling Society or the American Chemical Society's COMP and CINF divisions that facilitate the free flow of information, usually with a minimum of "hype". Our careers have been significantly shaped by interactions at such events which we consider to be a strong asset to the community that we should all try to support. One related action that perhaps deserves more attention is the capture of key learnings and experiences that we as a community can offer to the next generation(s) of scientists. Of course, the scientific literature must always be the first port of call, but are there ways to capture key "know-how"; hints, tips and anecdotes that may prove useful in the future? The emerging new media may offer a way to achieve this more effectively than the traditional route of refereed publication.

Weaknesses

In addition to external threats from changes in the pharma industry and from the relative youth of our community, computational chemistry also suffers from internal weaknesses within the discipline. In our view, the primary weakness is in the reliability of our underpinning theories and methodologies. There have been few, if any, significant developments in the core science in recent years. When one considers the spectrum of techniques in common use today, these are largely the same methods that were available in the early to mid 1990s. Moreover, while increased computational power enables calculations today to be performed faster or on more molecules, this improved computational capability has not necessarily led to any significant improvements in the overall efficiency of our processes. Hand in hand with this lack of progress in our core methodologies, much academic research attention has focused on a relatively small number of problems such as the prediction of free energies of binding, with much less effort (at least, judged by the literature) on less glamorous problems. We recognize that the problem of accurately and reliably predicting free energies of binding is the "holy grail" of our discipline, but the lack of progress towards a solution that is generally applicable suggests that some of this effort would be better deployed elsewhere. Moreover, the number of projects that lack a protein structure will continue to be substantial, despite the tremendous progress in structural biology in recent years. Thus, reliable structure-based calculation of binding free energies will not be possible for the majority of the problems we face in drug discovery. And even for those instances when one has sufficient structural data to attempt a prediction of binding free energy, it will rarely be the case that potency at the protein target will be the primary bottleneck in discovering a drug. Does this disparity reflect a disconnect between what are seen as the important problems in academia compared with those seen as important to computational chemists working in the pharmaceutical industry?

One view on what problems and methodologies are considered important in industry is provided by a recent survey of the computational chemistry community at GSK. The survey comprised just a single question, "What one or two or three programs would cause you to do me bodily harm if I were to take that software away from you?" In order to emphasize that we wanted to know general tasks and methodologies rather than specific versions of specific pieces of software, we provided a sample answer, "I need homology modeling software that allows me to simultaneously use multiple structural templates and that allows me to include ligands or co-factors during the model-building process." We got an 80% response rate to our survey, and we therefore believe that the results reflect a broad and accurate picture of the methodologies considered important to practicing computational chemists. The results of the survey are shown in Fig. 1; similar terms in the survey responses were consolidated, and the font size of terms in the word cloud is proportional to the number of respondents who mentioned that term.

A striking feature of the word cloud is that although molecular dynamics is widely represented in academic research programs and in the academic literature, molecular dynamics was listed as important by only one respondent in the GSK survey. In contrast, visualization of three-dimensional structures and visualization of experimental data are by far the tools most widely cited by survey

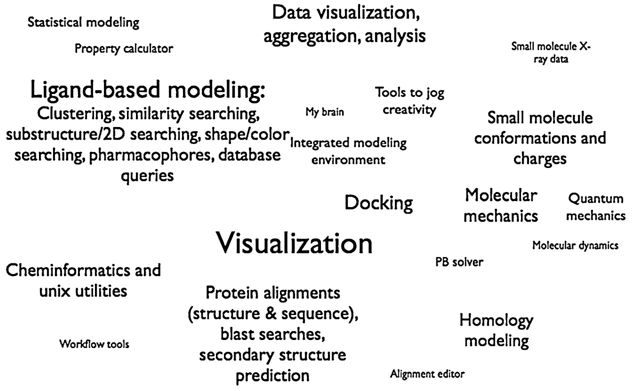
Statistical modeling
Property calculator
Data visualization, aggregation, analysis
Small molecule X-ray data
Ligand-based modeling: Clustering, similarity searching, substructure/2D searching, shape/color searching, pharmacophores, database queries
My brain
Tools to jog creativity
Integrated modeling environment
Small molecule conformations and charges
Docking
Molecular mechanics
Quantum mechanics
Visualization
PB solver
Molecular dynamics
Cheminformatics and unix utilities
Protein alignments (structure & sequence), blast searches, secondary structure prediction
Homology modeling
Workflow tools
Alignment editor

Fig. 1 Word cloud of responses from a survey of the GSK computational chemistry community

respondents. From an academic perspective, a general and reliable solution for computing binding free energies would mean that we as a discipline understand something deep about non-covalent association, and the study of such methodologies may therefore have fundamental research value. However, the word cloud in Fig. 1 suggests that there might also be fundamental benefit to health research if more resources were committed, for example, to better methods for reliably and rigorously extracting physically relevant patterns from three-dimensional structural data or from large experimental data sets.

In arguing that research effort might be profitably applied to areas beyond molecular dynamics and free energy simulations, we would also note that many computational chemistry methods appear to have reached a performance plateau. Even a cursory examination of the literature would reveal that the community has expended a huge amount of effort in evaluating methods for docking, scoring, virtual screening, pharmacophore elucidation, similarity searching and so on. Frustratingly, most of these studies confirm our assertion that our methods are not improving. Indeed, even when methods appear to be performing well for standard test cases, when evaluated in more realistic scenarios they demonstrate significant limitations. And yet we assert that despite the real limitations in our underlying methodology, computational chemistry still makes a significant impact on drug discovery. Much of the reason can be ascribed to the fact that very rarely are computational chemistry methods used without any intervention and guidance from a scientist often with significant knowledge of their target and often many years of experience in the application of different computational chemistry methods.

Strengths

We have argued above that the discipline of computational chemistry faces threats from the changing pharma landscape and from a dwindling of community memory. We have further argued that the underlying methodologies of computer-aided molecular design remain a weakness and remain unimproved over the lifetime of JCAMD and the lifetimes of the careers of the co-authors. We would further argue, however, that the strengths inherent in computational chemistry are numerous. A particular strength of computational chemistry in the pharmaceutical industry is demonstrated practical impact on drug discovery. As recognized in a recent article from the computational chemistry group at BMS [1], the question of "metrics" can be a contentious topic for "knowledge workers" such as computational chemists. While GSK does not use numerical metrics, the measures of success we use internally at GSK are very similar to those described by BMS. In preparing our annual accomplishments report, we record key contributions by computational chemistry to decision-making throughout the drug discovery process. Key impact statements from a recent year include the following types of contributions:

- Direct contributions to the discovery and design of two molecules that reached positive proof-of-concept clinical decisions
- Direct contributions to eight candidate and pre-candidate decisions
- On-going contributions to drug-discovery programs
    - Thirty seven highlighted contributions resulting in new hit/lead series
    - Eighteen highlighted examples of significant contributions to lead optimization
    - More than 70 examples of screening data analysis resulting in program progression
- Contributions to drug-discovery programs recognized in 25 published manuscripts and 12 issued or published patents

As can be seen from these impact statements, the true impact of individual contributions may only be apparent some time after the actual work is completed; synthesis timelines, patent considerations, and the amount of time required for molecules to progress through the clinic, all of these delay the evidence of impact across different time scales.

As computational scientists we are dependent upon many other disciplines such as medicinal chemistry, screening, structural biology, etc. to demonstrate the full impact of our work. This places a significant emphasis on our interactions with other colleagues and the personal relationships upon which those interactions are based. Put more crudely, having the greatest design idea is of no value at all if it cannot be converted into reality through dialogue and persuasion. To the extent that we are able to build strong relationships within the computational chemistry community and between this and other disciplines, that relationship-building skill is a significant strength that allows us to maximize the impact of our work on drug discovery.

And finally, our hardware tools have improved. With the possible exception of gene sequencers, we can think of no other discipline in the pharmaceutical research sector that has seen such an improvement in the performance of their instruments as have the computational sciences. Moore's law, first proposed in 1965, continues to apply and when price-performance is taken into account the improvement is even more dramatic. There is a strong argument that some developments in computational chemistry can be considered to have arisen directly from improvements in hardware, two examples being the rise of virtual screening of databases due to the inherent parallel nature of such tasks and the speed-ups now being achieved due to the use of GPUs [2]. Of course, computer hardware is useless without software. We have suggested above that the quality of the software which runs on these faster and newer hardware platforms has stagnated rather than improved over the lifetime of JCAMD. However, we continue to believe that increasing computational power will allow as a community to explore computationally costly and potentially more reliably predictive methodologies and will allow us to do the computational work necessary to put reliable error bars on the predictions of our current methodologies.

Opportunities

In the preceding sections, we have outlined the major threats to our discipline, have asserted that the poor reliability of our underlying methodologies is a weakness within our discipline, and have argued that the strengths of our discipline allow us to have real impact on drug discovery. In the final section of this short perspective, we will describe some of the opportunities that we believe will allow us to face the external threat of change in the industry, address the fundamental weaknesses in our disciple, and capitalize on our very real strengths to have even greater impact in the future.

First and foremost, computational chemistry needs to learn from the experimental sciences with regard to error estimation. In contrast to most experimental sciences, there is no real tradition in computational chemistry of providing an indication of the error of our calculations, with the possible exception of QSAR models. But even in the case of QSAR, authors of published models can appear ignorant of the fundamental fact that no model can be more accurate than the underlying data on which that model is built. A "good" biochemical assay or physicochemical measurement will typically have a standard deviation of around 0.3–0.5 log units when operated using a standard protocol in a single laboratory, and the corresponding value for literature data sets which are often derived from multiple laboratories will invariably be (much) higher. And yet, we often see reports of models that are "predictive" to significantly less than 0.3 log units. In addition to a lack of error estimation, a limitation of much of today's computational chemistry practice is that there is little analysis of the sensitivity of the results to changes in the initial state of the system or in values of input parameters. As we have suggested above, increased computational power makes it feasible to undertake such a sensitivity analysis by re-running a calculation using a number of different starting states and input parameters, and reporting the results of such variations. As a specific example, how many studies probe the variation of the charge and tautomeric state of active site protein residues, and the sensitivity of the predictions to these changes, in docking experiments?

The opportunity for the computational chemistry discipline is to emphasize a more technical and engineering approach rather than just a pure-science approach. Sensitivity analysis is one example of what should be common practice in computational chemistry, but is not the only such example. In the late 1990s, one of us had the privilege of collaborating with a world class Systems Engineering group, from which we learnt much about multi objective optimization and in particular Pareto methods. Engineering as a discipline has even more to teach computational chemistry. In the 1960s and 1970s, engineers defined and published well-characterized principles of engineering design [3]. Many computational chemists will argue that engineering design and drug design are substantially different, which is a valid argument but only to a point. Can any modeler disagree with this viewpoint:

> A designer must make decisions based on compromises involving innumerable factors—economic, human, technical and so on—and his final decision is one which optimizes all of these considerations. Unfortunately there is a widely held belief that decision-making is an art - an art practised by those willing to guess but unwilling to investigate alternatives considered rationally and simultaneously. There are those occasions when the scientific theories of decision-making through probability, statistics and optimization are ideally suited to the design method but a vital starting point is the quantification of all features of the compromise. At other times non-quantifiable factors and judgments of values render decision-making difficult but by no means insuperable. In fact mathematical methods can be used in these circumstances again by listing alternatives and isolating the most important variables.

Principles such as these can be directly applied to drug design: when designing a molecule to a protein active site, consider what effects the changes you are proposing to increase affinity might have on ADMET properties. Can you derive a predictive model which provides an estimate of the error (for example, using Bayesian or committee neural networks, or kernel methods instead of multiple linear regression)? Can we quantify the compromise between time taken to do a calculation versus the need to provide an answer to a particular level of accuracy and precision? As an experienced modeler, how can I quantify and document my insight and decision making, so that I can teach and mentor the next generation of modelers? Being able to articulate the Principles of Computational Drug Design—being able to work and publish in a more quantitative sense—would be a sign that our discipline is maturing to a level of "molecular engineering". In our opinion, we have some way yet to go to achieve this maturity.

In addition to the opportunity provided by application of an engineering approach to the practice of computational chemistry, we should explore the application of our techniques and expertise to new areas. The main focus to date of the computational chemistry covered by a journal such as JCAMD has been on the discovery of new chemical entities. However, our computational chemistry methods and skills can and should have application more broadly in important and growing areas of pharmaceutical research, biopharmaceuticals and pre-clinical development being but two examples. Discussions at GSK about opportunities in these and other areas have often led to fruitful collaborations involving the successful application of core computational chemistry techniques to important problems. Such areas may also present us with some additional "grand challenges" to consider, such as the prediction of crystal structures from first principles [4]. In addition to completely new problems, we believe that there continue to be considerable opportunities in the area of ADMET prediction, particularly as more corporate data is released into the public domain. Improvements in our ability to understand and predict ADMET properties may have an even greater impact on the overall success of drug discovery and development, given that attrition rates are still high.

Over the years, we have seen many new experimental techniques (e.g., high-throughput screening, combinatorial chemistry, full genome sequencing) introduced and often over-promoted as "the solution" to the pharma industry's problems. Invariably, over time each new technology finds its rightful place in the armory of techniques available to us in drug discovery. However, an oft-overlooked aspect of many of these new technologies is that they can present tremendous opportunities to those working in computer-aided molecular design. As a community we have well-developed analytical and problem-solving skills that are often critical to the success of new experimental methods. Two obvious examples would be the impact of design methods on the use of combinatorial and parallel synthesis methods and the development of approaches for the analysis and interpretation of large volumes of screening data. We are confident that whatever new technologies are developed there will be a need for our expertise to ensure their optimal implementation. These new experimental techniques invariably generate huge amounts of experimental data; better access to these data provides some key opportunities. The development of high-throughput assays and increased automation has enabled companies to generate large volumes of data on a diversity of chemotypes on physicochemical and ADMET-related properties such as lipophilicity, solubility, absorption, oral availability, hERG inhibition etc. Moreover, such data sets are no longer restricted to the industrial community with the availability of large databases of SAR data such as ChEMBL. Opportunities from large data sets include the increased use of data mining in order to derive new insights, or the use of corporate data sets as relevant and realistic blind-prediction data sets for theoreticians and software developers.

And finally, greater exchange of ideas and practices and people between industry and academia will provide new collaborative opportunities. The past few years have seen a much higher level of interaction between large pharma, smaller biotechs, and academia. In part this arises from the recognition that significant parts of drug discovery can be considered pre-competitive and that it is uneconomical for pharmaceutical companies to duplicate efforts. This higher level of interaction also reflects the funding provided in recent years for academic drug discovery [5]. Several of these projects rely directly or indirectly upon computational chemistry techniques. Three projects with which we are directly involved are the Innovative Medicines Initiative OPS (Open Pharmacological Space) and eTOX projects in the EU and the NIH-funded CSAR project in the US. The goal of the OPS project is to develop an open, public infrastructure for the integration of chemistry and biology data. eTOX will build improved models for toxicity prediction based upon data provided by the industry partners on failed development compounds. CSAR receives hitherto unpublished structures of protein–ligand complexes from industry labs and makes them available to the wider community in order to improve the performance of structure-based design methods such as docking and scoring. In all three of these examples there is a close and dynamic interaction between the industrial and academic groups around a common goal, more than has often been the case in the past.

Acknowledgments We found the SWOT approach to be an instructive and comprehensive way to assess not only computational chemistry at GSK but also the wider discipline. Our views may not all coincide with the views of others, nor may they be wholly comprehensive, but we believe that they represent a fair and balanced perspective. We would be interested to hear others' views on the strengths, weaknesses, opportunities and threats of contemporary computer-aided molecular design. We consider ourselves fortunate to have spent our professional careers working in a discipline that continues to grow and which we believe still has many exciting opportunities ahead. Finally, we would like to thank the members of the GSK computational chemistry community not only for their input to the analysis summarized above but also for their wider contributions as scientists and colleagues.

References

1. Loughney D, Claus BL, Johnson SR (2011) To measure is to know: an approach to CADD performance metrics. Drug Discov Today 16:548–554

2. Friedrichs MS, Eastman P, Vaidyanathan V, Houston M, LeGrand S, Beberg AL, Ensign DL, Bruns CM, Pande VS (2009) Accelerating Molecular Dynamic Simulation on Graphics Processing Units. J Comput Chem 30:864–872
3. Penny RK (1970) Principles of engineering design. Postgrad Med J 46:344–349
4. Kazantsev AV, Karamertzanis PG, Adjiman CS, Pantelides CC, Price SL, Galek PTA, Day GM, Cruz-Cabeza AJ (2011) Successful prediction of a model pharmaceutical in the fifth blind test of crystal structure prediction. Int J Pharm 418:168–178
5. Fry S, Crosby M, Edwards T, Juliano R (2011) From the analyst's couch: US academic drug discovery. Nat Rev Drug Discov 10: 409–410