

## A CADD-alog of strategies in pharma

Wendy A. Warr<sup>1</sup>

Received: 19 December 2016 / Accepted: 11 March 2017 / Published online: 18 March 2017  
© Springer International Publishing Switzerland 2017

**Abstract** A special issue on computer-aided drug design (CADD) strategies in pharma discusses how CADD groups in different environments work. Perspectives were collected from authors in 11 organizations: four big pharmaceutical companies, one major biotechnology company, one smaller biotech, one private pharmaceutical company, two contract research organizations (CROs), one university, and one that spans the breadth of big pharmaceutical companies and one smaller biotech.

The goal of this special issue on computer-aided drug design (CADD) strategies in pharma is to understand how CADD groups in different environments work. I have collected perspectives from authors in ten organizations: four big pharmaceutical companies, one major biotechnology company, one smaller biotech, one private pharmaceutical company, two contract research organizations (CROs), and one university.

Eric Manas and Darren Green of GSK emphasize “design” rather than “discovery”, since computational chemists are at their most effective when they apply true design principles. Since the process is a multi-objective one, and computational models have shown limited utility in predicting some essential parameters, Manas and Green believe that the industry is unlikely to be able to do CADD from first principles modeling and simulation alone, for the foreseeable future, and will need to use a mixture of theoretical and empirical approaches. They also emphasize the

importance of integrating the multiple sub-disciplines of computational chemistry, and discuss how resources are allocated to projects to maximize impact. Impact has been measured by a monthly reporting system of highlights; patent authorship; the level of repeat business; and allocation of extra staff resources. There are future opportunities in high performance computing; better force fields; cheminformatics techniques that automate SAR and predict reactions; and new statistical modeling techniques and optimization methods.

The contribution from Frank Brown and his colleagues at Merck also has an underlying theme of hypothesis-based design. The design group now includes computational chemistry, protein structure determination, and cheminformatics, in a global network, with a focus on delivering local impact. The authors describe the skillset and culture of the group, with an emphasis on leadership within teams. It is getting harder to recruit people with the right skillset; this has implications for the academics who will train the future generation. After the scientific and leadership engine is revitalized, external reputation is built, and this is seen in an increase in publications, and in attendance at meetings, sharing of ideas, and external collaborations.

The authors give examples of growing the footprint of the group into process and analytical chemistry. In discovery, impact is good if the team is successful at reaching conclusions faster, and in a more informed manner. In process chemistry, improvement in yield or enantiomeric excess might be a measure of successful modeling. The authors give examples of cost savings by reducing the need for counter-screening, and of success in analytical chemistry and multi-parameter optimization.

They also discuss in detail the choice of the best hardware and software environment. The approach is to maximize the value of the commercial and academic software

✉ Wendy A. Warr  
wendy@warr.com

<sup>1</sup> Wendy Warr & Associates, Holmes Chapel, Crewe, Cheshire, UK

portfolio, to turn methods into services, and best practices into workflows, and to broaden the availability of expert capabilities to a wider audience. The use of in-house GPUs or cloud computing is also discussed. The authors conclude by observing that all scientists need to improve their data science skills, since the most competitive companies will be the ones that use data best.

Herman van Vlijmen and co-authors discuss the activities of the global computational chemistry group within Janssen's R&D organization. A section on science covers bioinformatics and cheminformatics searches and data mining; data mining using linked data; ligand-based predictive modeling; virtual screening; quantum chemistry; and molecular dynamics and free energy perturbation calculations. The related software and databases (in-house and external) are listed. Collaborations are crucial to the operations of the group; several are discussed. It is not easy to quantify the impact of CADD, but one measure is the number of patents on which a CADD scientist is an inventor. The authors also discuss impact and value in a qualitative sense. In future there will be better predictions (including FEP), and improved machine learning methods (using, for example, descriptors from transcriptomics and high content cellular imaging readouts).

Ingo Muegge and co-authors describe CADD at Boehringer Ingelheim (BI). There are three core roles: (1) working closely with scientists in other disciplines, and sharing ideas, (2) turning data into hypotheses, to drive discovery and optimization of compounds, and (3) enabling medicinal chemists to use CADD tools on their own. The value added to a project by CADD is studied by assessing the satisfaction of the "customers" in the project team, by collecting feedback, and talking about mutual expectations.

A common platform shared by CADD scientists, structural biologists and medicinal chemists has been built around MOE. The authors describe predictive ADME modeling; matched molecular pairs analysis; and the BI Comprehensive Library of Accessible and Innovative Molecules (BICLAIM). CADD scientists also invest in computer-intensive technologies such as molecular dynamics simulations. They monitor trends, and evaluate new computational chemistry technologies. They have built a meta-layer which allows the front-ends used by medicinal chemists to be connected with the computational chemistry engines used in the back-end. Selected front-ends can be independently enabled to trigger meta-layer web services by using APIs or plugins. An increasing number of automatable CADD-related tasks are becoming amenable to use by medicinal chemists.

Many more experimental data, and sharing of pre-competitive data among companies will be needed in future. Cloud computing will encourage the development of more accurate, but computationally expensive methods.

Collaborations with academic groups will continue to play a key role, and crowdsourcing will provide access to a wealth of scientific talent. There is a growing need for CADD technology to be brought to bear on targets such as protein–protein interactions and RNA binding.

Jeff Blaney and his colleagues at Genentech discuss the goals and philosophy of the computational drug discovery group (which includes informatics and patent analysis, as well as typical computational chemistry); the project-centric environment, and group organization; the tools used; and the interdisciplinary skills needed. Performance "metrics" for both the computational drug discovery group and medicinal chemists are defined by the specific impact they achieved for the project that made a critical difference to the team. The computational chemist's work must lead to specific experiments: merely performing requested tasks is not sufficient. The authors describe in some detail the allocation of computational chemistry and software engineering resource to a project. Medicinal chemists, and other scientists, have been enabled to handle many of the more routine modeling and data analysis tasks. The interdisciplinary nature of the computational drug discovery group has implications for the education and training of future computational scientists.

Georgia McGaughey and Pat Walters describe the philosophy of the Modeling & Informatics group at Vertex Pharmaceuticals. The group reports to the Chief Scientific Officer. It consists of modelers, cheminformaticians and methods developers, from multiple disciplines (not just chemistry). It is not unusual to have more than one computational chemist working on the same project. The authors describe in detail the skills of the team and their project responsibilities. The impact of the group is measured through very regular performance reviews. The group is in the process of writing an infrastructure which will capture collaborative designs: a move toward quantitative metrics.

The Vertex integrated informatics infrastructure was built internally but relies heavily on software components from OpenEye and ChemAxon. The authors describe features of the ELN; the modeling and design tool; and the searchable document repository. There is a large set of scripts implemented in Python. Both commercial and open source tools are used.

Integration of internal information with that from external public and proprietary databases is important. More sophisticated ways of visualizing complex read-outs are becoming routine in the dissemination of knowledge. Pre-competitive knowledge sharing is happening; for example, ten pharmaceutical companies are comparing experience with free energy methods. In future, there will be tighter integration of data from proteomics, biology, chemistry and computational science. Computational chemists are also moving into areas such as biocatalysis, toxicology,

polymorph prediction, and process chemistry, with increasing predictive power.

Brock Luty of Dart NeuroScience and Peter Rose of the RCSB Protein Data Bank argue the need for specialist scientific software engineers. These engineers are essential in creating a software foundation that is maintainable, validated and robust. Research informatics software engineers do not always have a deep background in science, and computational chemists do not always have the time or skills to do proper software engineering. Scientific software engineers bridge the gap.

Steve St-Gallay and Colin Sambrook-Smith of Sygnature Discovery present a CRO's perspective on goals; the skills, software and hardware needed; and the approaches taken to design quality compounds for synthesis. They justify their arguments that free-energy perturbation techniques provide only questionable value to medicinal chemistry programs at present; that use of Amazon Web Services is a flexible and secure option; and that applying metrics or business process improvement techniques might in fact have a negative effect. Their most compelling evidence of impact is that customers keep on coming, with a significant amount of CADD effort often specified in the contract. The authors think that OpenEye's Orion could be very significant in future, especially since it will reinforce one of their strengths: communication with their customers. They are also investigating virtual reality using Oculus Rift and Molecular Rift.

Douglas Kitchen of Albany Molecular Research, Inc. (AMRI), another CRO, discusses the contributions of computational chemistry and cheminformatics in chemical library design, hit triage, hit-to-lead, lead optimization, and structure-aided design, and the techniques used. The hardware system used is effectively a private cloud solution. Kitchen conveniently tabulates all the software used and its functions; AMRI is making increasing use of open source software and public databases. For many reasons it is difficult to quantify the contribution of CADD to success in drug discovery programs. Issues include the collaborative nature of the process, the multidisciplinary nature of the teams, and the different expectations of calculation accuracy. In justifying the impact of CADD, Kitchen gives detailed scientific examples. He concludes with a

discussion of future opportunities including the demands posed on computational chemistry by some new experimental techniques (e.g., image analysis, phenotypic screening, and techniques to measure binding).

The paper by Gerhard Ecker's team at the University of Vienna is centered on linked life science data. Public availability of sources such as ChEMBL and the Open PHACTS Discovery platform, and use of the data with workflow engines such as KNIME and Pipeline Pilot, have allowed the team to expand its approaches from conventional Hansch analysis to complex, integrated multilayer models. Using open source tools, predictive models can be built inside the data curation workflows, or as a separate instance. A recent case study exemplifies the seven pillars of the process: (1) collecting relevant pharmacological data, (2) filtering them from artifacts, (3) building computational models for prediction of novel compounds, (4) selecting the best model, (5) prioritizing virtual hits for experimental testing, (6) using experimental ligand-based data to guide molecular docking studies, and (7) identifying plausible binding modes. The method allows challenging targets to be investigated. Ecker and his co-authors also illustrate the significance of consortia and precompetitive alliances.

Terry Stouch and I presented the invited authors with a large list of questions and topics. Almost all of the topics are addressed by one or more of the papers. Outsourcing of CADD was a notable exception. One additional topic that appeared twice was the need for changes in the education of the next generation of computational chemists. Have I learned anything new? Probably not: I expected to read about the changes that have happened over the last decade (an end to the separation of computational chemistry, cheminformatics and information science; the rise of data science; integration of internal and external databases; open source; cloud computing; routine use of CADD tools by non-experts, FEP, involvement in new technologies and disciplines, etc.). Nevertheless I do feel that these eleven papers make an interesting and useful compilation.