Organic and Medicinal Chemistry
International Journal
ISSN 2474-7610

Juniper Publishers
key to the Researchers

Opinion
Volume 11 Issue 1 - July 2021
DOI: 10.19080/OMCIJ.2020.11.555801

Organic & Medicinal Chem IJ

Predictive Modelling, an Opportunity for the Chemist, a Challenge for Method and Software Developers, and a Task for Science Journals

Robert J Meier*

Pro-Deo Consultant, North-Rhine Westphalia, Germany

Submission: June 17, 2021; **Published:** July 05, 2021

***Corresponding author:** Robert J Meier, Pro-Deo Consultant, 52525-Heinsberg, North-Rhine Westphalia, Germany

Opinion

Modelling has become widespread in the science community, practiced by both the experts and non-experts, also in organic and medicinal chemistry. Tools have, in part, become available in user-friendly form such that the organic chemist, and any scientist, can perform molecular dynamics simulation, quantum chemical calculations or data driven (models parameterized using available experimental data) predictive modelling. However, as for performing proper experimental work it generally requires expertise, i.e., an understanding of the science behind to arrive at reliable and justified results. When we focus on property modelling, where properties include boiling point, melting point, water-octanol solubility coefficient, toxicity and so on, if an organic chemist synthesizes a molecule for a purpose it will need to have certain properties. Modelling tools could help the chemist to select a molecule with the desired properties, or at least part of it. That there is great potential in such modelling tools can be recognized realizing that there are more than 1 billion organic molecules with 13 heavy (non-hydrogen) atoms (the GDB-13 database [1], whereas GDB itself is just a data base file format), and because of the up to 13 heavy atoms this is only a subset of all possible organic molecules. When we need access to properties of any molecule, namely the desired molecule or a molecule whatever size but with certain properties, it will be evident that determining these properties experimentally and then selecting the correct molecule is undoable. Thus, it would be very advantageous if reliable predictive tools could be used by the relatively non-expert user as this could greatly facilitate the work of the, e.g., organic, chemist. The opportunity is to be benefit of the experimentalist, but there is a challenge for developers of the tools and software as we will argue below. Also, journals could be more critical in accepting papers in which methods or data employed are not made fully public.

The user

There are expert and non-expert users. The expert users can be divided in the group of users that were close to the development, e.g., all ex-students from an academic group developing models, or the employees of a software company, and the group of experts in theory and modelling tools but not having had direct access to all available knowledge. For the non-expert users, we can divide between critical users that might and go for help from experts, and those that do not. The first group of all four groups is likely to produce proper results without any exception, but this will generally be a very small group. The last group will be most vulnerable to producing improper results. But, whatever the level of expertise the user has, it is detrimental by any means if there are serious flaws, or what are called bugs, in the software, which does happen unfortunately as we will see further below.

These is, however, another relevant item, namely what is required from a software tool for the problem the user wants to handle. For some cases we need a very accurate and reliable answer, for other problems we need a ball-park figure. If we need a homogeneous catalyst with high selectivity, we accept a model that predicts the upper 10% (so selectivity between 90 and 100%) correctly while we have a lot of scatters in the lower range which is totally irrelevant as these are not the potentially interesting catalysts. The same applies to issue in the field of toxicology: according to current practice in that field a substance is, e.g., carcinogenic, or not. This means only those predictions close to the borderline 'carcinogenic vs. non-carcinogenic' should highly accurate. It depends on the target which quality the method should have to serve its purpose for that target. This is often not realized and not discussed in papers. Results are often praised, without a proper discussion on the true requirements for a useful

and reliable answer. A more extensive account on this theme was presented in [2].

In many cases the non-expert user might select the model or tool which has been reported as better than what one had before. However, method developers often report the quality of their tools by quoting a quantity such as absolute mean deviation (from experimental values). But that an averaged deviation is smaller than before does not mean the method is a priori better. It could be for instance that a few selected cases have a very substantial deviation, but as we do not know to which molecules this applies, it makes such a method inappropriate (unreliable) to be taken as the basis for decisions. The averaged deviation should be small, but at the same time the maximum deviation occurring at all should be within a certain well-defined range depending on the problem. For example, when we discuss reaction energies and need individual heats of formation, a small average deviation does not lead to a sufficiently reliable value, i.e., within 1 kcal/mole ('chemical accuracy') when individual values might be far off. What the experimentalist needs is a method which leads to reliable predictions that lead to the correct choices with all individual predictions within 1 kcal/mole as in [3].

Software Issues

Some software packages are cheap so that every individual can afford them, so to speak. But there is also a lot of freeware. Other software can be really very expensive. But more important than that is the fact that major errors have been found in software, also in very well-known software suites, leading to erroneous results where it was in almost impossible for most users to be aware of this. When such errors pop up late, when many publications are already out, we cannot judge the reliability of some of the results in retrospect unless we repeat all the work with the correct code. These are issues which are generally not discussed in the 'normal' scientific journals, reporting negative results is not done (why not? as it can contribute as much as positive results: you learn from your mistakes, not from all those things that go right straight away). A year ago, there was an interesting and still very worthwhile to read article in Chemistry World, the Magazine of the Royal Society of Chemistry, entitled 'Computational chemistry faces a coding crises' [4]. We cite 'This is not the first time that an error in a piece of software code has cast a shadow over computational research, these sorts of issues are actually surprisingly common. In one famous case, a coding error was at the heart of a seven-year dispute between some of the world's top theoretical chemists, who were trying to model the phases of supercooled water.

And recently, an algorithm used in older versions of the popular molecular dynamics software Gromacs was found to introduce order of magnitude mistakes during simulations.' Furthermore; 'Ideally, code will be well documented and publicly available, allowing researchers to scrutinize scripts and locate problems. But this is not always the case – traditional publishing practices, as well as concerns around intellectual property, often mean that code is difficult or even impossible to access '. These seem a few cases, but from my 40 years' experience as an experimentalist and computational scientist, I have experienced such situations more often. Only in-depth knowledge of the science behind, a critical mind and in particular testing the software to simple well-known cases (where the answer is unambiguous and known) has saved me from publishing erroneous results. But obviously testing a software tool by reproducing published results with the same software is no guarantee at all. Thus, by experience, there is a very important task for software developers to ensure their codes are fully correct and reliable, the commercial ones as one has spent (a lot of) money on this. Of course, there are codes which are fully reliable, but unfortunately that does not account for all.

Method Development

It is not always only the software itself, but it also happens those methods are not fully published. For some data driven models the data set employed to parametrize the model has not been made public. This means the results are not verifiable independently, and secondly the user cannot see to which molecules or otherwise other entities the model was parametrized, and thereby the domain of applicability. Journals should not accept papers which do not contain the full details. Authors might not want to make all data available, e.g., when there is commercial interest, but then a manuscript should not be regarded as suitable for publication in a scientific journal as scientific results should be verifiable. This is a basic element in science, and it should be acknowledged for the full 100% (the author has worked in industry over 3 decades and is thus familiar with the issue). This also touches the well-known and long-standing issue of validation. This has been an issue for decades in the field of molecular modelling. Also, for commercial and expensive software packages this has more often been a critical issue. It may not be in the interest of software vendors to acknowledge that the tools do not have an 'infinite' domain of applicability, but it is crucial to the user to have this knowledge, or at least the awareness.

Is there light at the end of the tunnel? Of course most of the work published will be fine. But for part of work, it is difficult to verify. In principle the solution of most issues is relatively straightforward, and in many cases it 'only' requires some additional work and discipline. Full public availability of the details of the method, and journals not accepting manuscripts that do not contain such information should not accept manuscripts not providing this. Good references to the methods, a good manual accompanied by descriptions of limitations, and validation studies (either in the manual or in separate, available, scientific papers or otherwise accessible documents) must be considered an absolute need. Currently we see what is almost a hype, namely Artificial Intelligence applied to all problems you can imagine. It would be great if here the issues addressed would be accounted for properly right from the beginning.

How to cite this article: Robert J M. Predictive Modelling, an Opportunity for the Chemist, a Challenge for Method and Software Developers, and a Task for Science Journals. Organic & Medicinal Chem IJ. 2021; 11(1): 555801. DOI: 10.19080/OMCIJ.2021.11.555801

References

1. Fink T, Reymond J L (2007) Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: Assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes and drug discovery. J Chem Inf Model 47: 342-353.
2. Meier R J (2019) A Way towards Reliable Predictive Methods for the Prediction of Physicochemical Properties of Chemicals Using the Group Contribution and other Methods. Appl Sci 9: 1700.
3. Meier R J (2021) Group contribution revisited: the enthalpy of formation of organic compounds with "chemical accuracy", Chem Engineering 5: 24.
4. https://www.chemistryworld.com/news/chemistrys-reproducibility-crisis-that-youve-probably-never-heard-of/4011693.article

DOI: 10.19080/OMCIJ.2021.11.555801

How to cite this article: Robert J M. Predictive Modelling, an Opportunity for the Chemist, a Challenge for Method and Software Developers, and a Task for Science Journals. Organic & Medicinal Chem IJ. 2021; 11(1): 555801. DOI: 10.19080/OMCIJ.2021.11.555801