The cost of prediction. How computational methods compromise reproducibility

Johannes Lenhard, Simon Stephan, and Hans Hasse

Philosophy in Science and Engineering, RPTU, Kaiserslautern, Germany

Accepted as a chapter in "The Perils and Promises of Prediction", edited by Th. Arabatzis, S. Arapostathis, I. Katsaloulis, and A. Tympas, to appear in 2024.

Abstract

This paper examines a looming reproducibility crisis in the core of the hard sciences. Namely, it concentrates on molecular modeling and simulation (MMS), a family of methods that predict properties of substances through computing interactions on a molecular level and that is widely popular in physics, chemistry, materials science, and engineering. The paper argues that in order to make quantitative predictions, sophisticated models are needed which have to be evaluated with complex simulation procedures that amalgamate theoretical, technological, and social factors – leading to problems with reproducibility. Thus, for methodological reasons, the predictive success causes a reproducibility problem.

1 Introduction

That we live in the age of science is one of the most agreed upon views of what is characterizing our times. Scientific knowledge is widely appreciated in society and the perceived ongoing progress in science persistently raises the bar of expectations. Prediction is the currency in which science fulfills (or has to fulfill) the expectations. Forecasting the weather, designing technology, giving policy advice—all sorts of application depend on some predictive capacity. What is able to make predictions has got something right and is therefore held in high esteem.[1] It is all the more annoying when a claim for prediction fails where it was

[1] Making this rough sketch of an argument more precise leads to deep and ongoing philosophical debates. Prediction somehow is a hard currency, but it does not directly translate into truth of a theory or hypothesis. Challenging philosophical issues abound: What exactly follows from accurate prediction, if not truth? Furthermore: how can one explain the success of a theory that is not true? In what sense is the world welcoming to simplified and

taken for granted—and science falls short of expectations. This is the case when even the seemingly known outcome of an experiment that has been carried out before cannot be predicted. A recent example is the reproducibility crisis in medicine and psychology: when targeted studies failed to reproduce a large fraction of published studies, this provoked an uproar in the scientific community and the media. A common opinion is that good science can show high predictive capability, whereas a science that cannot even reproduce its reported results is not science at all. In a highly shortened form: non-R (not reproducible) implies non-P (no serious prediction).

This paper tells a very different story: It examines the cost of prediction, i.e., how those methods that produce valuable predictions have potential (unwanted) effects. Different from most accounts of the reproducibility crisis, which target psychology, medicine, and sociology, this paper probes the core of the hard sciences. Namely, it concentrates on molecular modeling and simulation (MMS), a family of methods that, very roughly described, predict properties of substances through computing interactions on a molecular level and that is widely popular in physics, chemistry, materials science, and engineering.[2] Put in a nutshell: Even the champions of prediction are experiencing problems with reproducibility. The fundament is crumbling. Furthermore, it is argued, for methodological reasons, the predictive success causes a reproducibility problem. Thus, P implies non-R.

The claim appears to be hardly acceptable because it indicates a contradiction: P -> non-R -> non-P. This paper explains why this assertion is nevertheless reasonable—the logical formulation is misleadingly concise. Essentially, molecular simulation cannot maximize

---

idealized accounts? These questions are often (not always) discussed under the rubric of (scientific) realism (see, for a recent example, “the argument from successful theories” in Patton 2023, section 4.1 and the literature she cites there.) We will briefly touch upon this issue in the concluding section.

[2] This popularity is based not only on the feasibility of making predictions, but also on the match with empirically measured data. In general, such agreement provides reasons for accepting simulations as a method for making valid predictions (see, e.g. Lenhard 2019). Kampouridis (2022) shows the importance of this comparison in the case of quantum chemistry. Notably, although simulations match measured data with high precision, this is not tantamount to full reproducibility.

predictive power and, at the same time, attain full reproducibility. The paper argues that in order to make quantitative predictions, sophisticated models are needed which have to be evaluated with complex simulation procedures that amalgamate theoretical, technological, and social factors – leading to problems with reproducibility.

Section 2 distinguishes two viewpoints regarding a lack in reproducibility: Is it a crisis because something is going seriously wrong in science? Or does such shortcoming merely indicate a scientific problem, i.e., something that science normally addresses and solves?[3] In molecular simulation, the problem is acknowledged, but not perceived as a crisis – just something to be solved. Based on recent literature in the field, section 2 discerns two different strategies for obtaining a solution. One is "forcing" reproducibility from below, the other "smoothing" the problem from above. It is argued that these existing strategies dissolve the problem rather than solve it. In general, this is not a bad thing. However, we claim, only a third strategy can address the *reasons* for the problem with reproducibility. Such third strategy must be based on a detailed examination of the process of simulation modeling.

Section 3 contributes to this examination and presents the core of our argument in three steps. The section first introduces molecular dynamics (MD), a principal member in the family of MMS. MD is a computational method that builds on Newtonian mechanics and statistical thermodynamics. It models the interaction between particles and employs the computer to calculate the resulting interactions between very many particles, thus obtaining predictions of how substances behave. The argument does not strongly depend on choosing MD; any other member of MMS (like Monte Carlo) would yield similar results. Moreover, the case arguably generalizes to a much larger area of simulation modeling. One big advantage of MD is that one can discuss the case while largely avoiding technical apparatus. The second step reports from a study (Schappals et al. 2017) that indicates gaps in reproducibility. This study did not follow one of the solution strategies, forcing or smoothing, but rather asked: to what extent do simulation experiments yield reproducible results when the *same* theoretical model is used by different groups that work at different locations with different software implementations that run on different computers? This study design invites to analyze the steps between a (given and fixed) theoretical model and the simulation results. Nobody would expect that the simulations come out identically, but they diverge in an interesting way beyond statistical

[3] According to Thomas Kuhn (1970), problem (or puzzle) solving is the way in which normal science makes progress.

uncertainty, and thus invite an examination into the reasons for larger-than-expected non-reproducibility.

The third step undertakes this examination. Here, we connect to and build on existing literature in the philosophy of simulation modeling that has studied how simulation modeling extends and also transforms older conceptions of theoretical and mathematical modeling (Humphreys 2004, Winsberg 2010, Gramelsberger 2011, Lenhard 2019, among others). One important methodological twist that makes our approach feasible is the close cooperation between simulation practitioners and philosophers. The upshot is that further unpacking the modeling process is not (or need not) getting lost in technical detail, but rather yields new insights into how epistemological, methodological, and social factors interact in the problem of reproducibility.

The concluding section 4 argues for the claim that the problem with reproducibility points toward a crisis, namely an identity crisis of simulation models. According to the normal standpoint, the identity of a simulation model is founded in the theoretical model.[4] However, the analysis in section 3 shows how predictions result are influenced by a plethora of intertwined factors. In effect, the predictions arise from how theory, computational technology, and a host of variable factors in computational processing work together. Moreover, it is hardly possible to trace the role of these factors separately. Thus, the identity of a simulation model seems to have a somewhat disconcerting—and philosophically challenging—holistic character.

Finally, we take a look beyond our case study and toward machine learning with deep neural networks. There, parameterization is taken to the extreme, maximizing predictive capacity. In line with our study of molecular modeling and simulation, the cost of prediction seems to surface in a repeatedly observed non-robustness of the predictions.

2 Crisis or (normal) problem?

[4] We leave aside cases where a theoretical model does not exist, or at least does not play an important role, like in many agent-based simulations that model actions and interactions of agents without starting from a more general theoretical description (Epstein and Axtell 1996, Wise 2004).

Reproducibility is firmly accepted as a basic condition for science. From a historical perspective, it is a hard-won condition. It comprises an institutional component, because the very point of reproducibility is that it does not hinge on single gifted persons, rather is of a systematic and communal character. Different scientists at different locations, pursuing the same experiments get the same results.[5] What counts as the "same" experiments is not obvious. Neither is what counts as the "same" results. They usually are not identical, but coincide in the sense of some statistical measure. In general, evidence counts as scientific only when it is reproducible. By implication, if it is not reproducible, it is not scientific. This rule is not as strict as it looks from the outset, because whether something is reproducible—or can be made so with a bit more time and effort—can be a delicate question. Consequently, a failure in reproducibility may be perceived either as a challenge, i.e. a problem to be solved, or as a crisis.

The reproducibility (or: replication)[6] crisis arrived with a splash in the (scientific as well as popular) media in the 2010s when an increasing number of meta-scientific studies indicated that reported results from a number of scientific studies could not be replicated.[7] Focusing on psychology and medicine, one prominent contribution is *Open Science Collaboration* (2015) where a large number of expert groups made an orchestrated effort to replicate over hundred psychological studies—and was successful only in 39%. What followed were pleas to the ethics of researchers, but also to "Open Science".[8] In short, respected scientific fields that

---

[5] Moreover, a goal can also be to confirm a result or phenomenon by different experiments.

[6] The terminology (reproducibility – replicability – repeatability) is infamously confused. Different scientific disciplines follow different conventions, cf. Pleisser 2018, NASEM 2019. To avoid the discussion about what term is appropriate in what situation, we stick to reproducibility and use it in a fairly generic fashion, specifying the meaning where necessary.

[7] A much-cited pioneer is John Ioannidis with his 2005 essay "Why most published research findings are false." The tag as a "crisis" was added by Pashler and Harris 2012. The replication crisis even has an entry in wikipedia.

[8] Atmanspacher and Maasen edited the volume *Reproducibility. Principles, Problems, Practices, and Prospect* (2016) that discusses a range of related topics from a sociological and science studies point of view. The report *Reproducibility and Replicability in Science*

emulated the research methodology and publishing policies of the "hard" sciences, faced unanticipated problems. The ensuing debate about the reasons that led to the lack in reproducibility identified a number of (possibly) contributing factors. Sometimes the publication had not fully accounted for the actual complexity of the study, various sorts of biases might have played a role, or even outright fraud. Or even more distressing, the crisis might teach a lesson not about mistakes, but rather about the regular quality and trustworthiness of scientific findings.

However, the common perception of the crisis is that fields, which aspire to be on a par with the "hard" sciences, fail to live up to their claim. Consequently, the predictive capabilities of those fields involved in the crisis are called into doubt—whereas both reproducibility and predictive capacity of the hard sciences remain unquestioned. This does not mean there would be no problems with reproducibility.[9] Such problems exist and are acknowledged, but they count as normal problems to be addressed, not as crisis—as, for instance, the NASEM (2019) report of the American Academies illustrate.

Many scientific fields have recognized that reproducibility emerges as a new problem in the context of computational methods. In this regard, molecular simulation is a typical case. From here onward, we focus on this case. Two aspects feature prominently in the discussion about ongoing digitization. The first issue is the status of data. Data about properties of materials are a typical and central part of scientific results. Such data are gathered in large data bases. New data from experiments (that normally involve computational models, too) are checked whether they fit to existing data, in the sense of reproducing these data in a statistically acceptable way (cf. Cummings et al. 2009, Brennecke et al. 2019). The second issue is about formulating conditions so that simulated data can claim to be (at least) on a par with experimental data (Kofke 2016).

A third aspect is currently gaining momentum. It is concerned with the side of modeling and simulation. There, the reproducibility problem is whether *simulations themselves* are reproducible. The current paper focuses on this problem. In the literature, one can discern two

---

(NASEM 2019) documents that the issue has arrived in the highest echelons of scientific institutions.

[9] One can think of high energy physics where there might exist only one laboratory that is able to conduct certain experiments. These experiments are seen as reproducible in principle, whereas in practice, they are not.

strategies for solving the reproducibility problem. Lejaeghere et al. (2016, 2020), Mueser et al. (2017), or Wan et al. (2021) approach the problem "from above". They define test cases and then organize large community efforts that study these cases through extensive simulation. Participating groups use those models and software they are specialized in. The overall study then observes whether the various results give a coherent picture in the limits of some statistics. This approach can paint an overall picture where the various simulations deliver predictions that are reasonably close to each other. In other words, this approach documents to what extent simulations reproduce other simulations (in the test case and in statistical limits), without taking care of different modeling assumptions and implementations on a finer scale. This approach can signal a green light, but cannot provide reasons for where reproducibility is getting into trouble.[10] Thus, the strategy can be called "smoothing from above".

A second solution strategy approaches the problem from the opposite side, forcing "from below". This strategy is looking for an institutional solution, a standardization of methods that prescribes the way simulation studies should be done so that they can be reproduced. The proposed measures include publishing the code (open science, transparency), and tying simulations to a controlled software environment, thus unifying the proliferation of current models, softwares, and computers. These measures approach the problem from below and try to nail down all conditions in sufficient detail so that simulation studies can implement all conditions of an earlier study and then reproduce the results of this study. A considerable movement has emerged that contributes to this strategy (Thompson et al. 2020, Abraham et al. 2019, Horsch et al. 2020, Gygli and Pleiss 2020). The goal is to reproduce the very same simulation experiment, or to make a published simulation study reproducible in principle. Hence the problem of reproducibility is seen from a somewhat narrow perspective, forcing reproducibility from below, whereas the question what happens when *different* simulations

---

[10] The big model comparison projects in climate science (a central part of the IPCC reports) take place in a similar situation. Different institutions and centers contribute extensive studies, all in a specified test bed, but each center using its own circulation model. In this way, one can diagnose the extent of agreement, which is important. But if differences occur, it proves hard or impossible to attribute the differences to particular causes/ modeling assumptions (cf. Winsberg and Lenhard 2010). In this setting, reproducibility means that different approaches to the same task are in mutual agreement.

evaluate (in the technical sense of assigning the numbers under specific conditions) the same theoretical model is slipping out of sight.

Both strategies dissolve the problem rather than they solve it, because they either find that there is no problem (strategy 1), or they avoid the problem (strategy 2). In general, this is not a bad thing (the problem is gone). However, in our particular case, insight into the epistemology of modeling would be suppressed. Neither of the two strategies can serve the claim of our paper, i.e., that increasing predictive capacity is a reason for problems with reproducibility. The argumentation for the claim requires a different setting. The guiding question is to which extent simulation experiments produce reproducible results when the same theoretical model is simulated at different locations (institutions) from different groups through different software implementations that run on different machines. The variations that matter are those that are caused by modeling steps in between the theoretical model and the simulation result. Only on this level can one identify reasons for problems with reproducibility and find arguments about the limitations of reproducibility.

3 Our case: molecular dynamics and a round robin study

In many fields of science and engineering, simulation modeling starts from a theoretical mathematically formulated model. The latter is then said to be evaluated by simulation experiments, i.e., values assigned through simulation runs under specific conditions. These experiments "live" on simulation models. Many practitioners assume that these simulation models give an accurate picture of their theoretical starting point (in the limits of controlled approximation and statistics).[11] We take a practice-oriented stance: Whether the simulations give in fact accurately picture their theoretical model is a question not of philosophical or technical definitions, but rather of actual scientific practice.

[11] Work on the epistemology of simulation repeatedly has sent warning signals. This work has brought to the fore that the relationship between theoretical model and simulation is a complicated, though philosophically highly interesting one, cf. the monographs Humphreys 2004, Winsberg 2010, Weisberg 2013, Morrison 2015, Lenhard 2019 that take different, though related stances.

Our analysis proceeds in three steps. Firstly, we briefly introduce Molecular Dynamics (MD), which is a simulation technique that investigates properties of materials[12] by the following straightforward recipe. Model the interaction of particles via classical mechanics and observe the time evolution of the system, then extract properties of interest from these simulations. MD simulations numerically solve the Newtonian equations of motion simultaneously for all particles. The scope and precision of predictions made MD a popular tool in science and engineering. Secondly, we discuss results from a round robin study that assigned the task of simulating one and the same model to different expert groups, working at different locations and with their own implementations (Schappals et al. 2017). This study reports problems with reproducibility that were not anticipated by the practitioners and that pose a serious challenge. Thirdly, we analyze the factors that contribute to this problem. The main suspect is an over-simplified picture of the process that leads from the mathematical model to the simulation outcome. Only through analyzing all modeling steps that lead from the theoretical model to the concrete implementation, can one find out the reasons for the reproducibility limits as well as their delineation.

3.1 A Primer to Molecular Dynamics

Molecular modeling is based on the fundamental idea that macroscopic behavior of matter results from the interaction between small particles. Since Greek atomism, this idea found varying expressions. However, it was a way for explaining observed material properties, whereas prediction remained out of the question.

Such prediction became feasible in the process of mathematization and mechanization that shaped modern science (Dijksterhuis 1961). After Newton's theory of gravitation had provided an example of striking predictive capacity (based on particles and forces), Laplace framed a program according to which this approach should be emulated in all of physics (Fox 1990).[13] Laplace is famous for refining the mathematical apparatus for celestial mechanics. And he famously had high expectations. According to him, the physical dynamics of the

---

[12] In addition, MD can also be used to study nanoscopic processes based on the knowledge of the properties of the materials - an application that we have not included in our discussion, as it would not change the picture.

[13] Obviously, the historical process cannot even roughly be captured by jumping from hero to hero over centuries.

universe can be fully predicted if only one knows the acting laws between particles, the starting conditions and one is able to solve (integrate) the Newtonian equations for all particles of the universe simultaneously. Of course, he was fully aware that the necessary mathematical capabilities were beyond the reach of human beings, rather called for a superhuman power, the later so-called Laplacian demon. Nevertheless, prediction, even a dream of perfect prediction, had become a topic that could be discussed in the context of a mathematical-scientific method.

We jump forward again. With the computer as instrument, molecular modeling and simulation became feasible.[14] MMS is a family of methods that combines Newtonian mechanics with statistical physics and computational methods. Molecular Dynamics (MD) is arguably that member of MMS with the closest kinship to the Laplacian idea. The basic recipe is simple: one models the interaction of the particles via classical forces and then computes the resulting behavior by numerically solving a large number of differential equations. In a way, MD employs the computer to emulate the Laplacian demon.[15]

The recipe is simple, its execution is not. We provide a sufficient feeling for the (relevant and non-technical) intricacies by describing the first step, namely modeling the most basic component of interaction. The go-to real substance is the noble gas argon. The atoms of argon are spherical and the only relevant forces between them are those resulting from repulsion and

---

[14] In the early 20th century, quantum theory made it clear that the interaction of very small entities, like electrons forming a bond, cannot be described by classical forces. However, the quantum theoretical treatment of systems with many molecules remains largely intractable even with the computational power that is available today. Molecular modeling occupies the space in between (sub-)atomistic quantum mechanics and continuum mechanics where the discrete nature of the molecules can be neglected. On prediction and computation in quantum chemistry, see Kampouridis's chapter in this volume and also Johnson and Lenhard (2024, chapter 4).

[15] For an early history from a practitioners' perspective, see Battimelli et al. (2020). Rowlinson (2004) covers the long-term perspective on "coherence", starting from Laplace, with a wealth of scientific literature.

dispersive attraction.[16] In principle, all argon atoms in a many-particle system interact, but computing this turns out to be basically infeasible. A common simplification is to assume that the interactions in the system can be represented by pair-interactions, i.e. that it is sufficient to consider only interactions between two partners (which are then assumed to be independent of what the other atoms do).

However, what is the adequate mathematical form of the pair potential?[17] Finding suitable forms is far from trivial, even for the simple example of argon, the ansatz can be formulated in various ways; but all of them contain parameters that have to be fitted to data. The most popular ansatz for doing this is the Lennard-Jones potential, named after the pioneer of quantum chemistry, Sir Lennard-Jones (1894-1954).[18] It is given in Equation (1) that we display because readers can grasp how it serves the argument without needing any expertise for actually handling such expressions. This potential consists in the superposition of two exponential terms, the one (with the exponent *m*) controls how quickly the repulsive force rises when bringing two particles closely together, the other term (with the exponent *n*) expresses how quickly the attracting force decreases with increasing distance between the particles (*r* denotes the distance between them).

$$V(r) = 4\epsilon[(\frac{\sigma}{r})^m - (\frac{\sigma}{r})^n] \qquad (1)$$

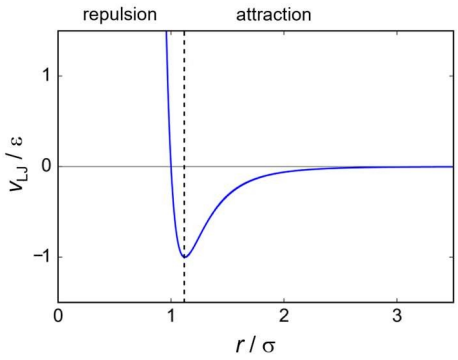
Lennard-Jones proposed the exponents $m = 12$ for the repulsive term and $n = 6$ for the attractive term as adequate choices (Lennard-Jones 1931). Basically, he chose $n = 6$ because

---

[16] This attractive force is also called the van der Waals force. By the way, Primo Levi begins his celebrated *The Periodic Table* (1975) with equating his family with argon because of their reluctant and weak interaction with the outside world.

[17] Usually, the potential energy is modeled and the force is obtained by derivation. Hence the pair potential gives the force acting between these pairs.

[18] For a historical study of philosophical transformations linked to methods of computation along the case of the Lennard-Jones potential, see Lenhard, Stephan, and Hasse (in preparation).



Fritz London (1930) had calculated this exponent from quantum theoretical considerations.[19] More precisely, London had examined hydrogen and calculated from the Schrödinger equation that the van der Waals force between two atoms decays with the sixth power of the distance between the atoms. After setting $n = 6$, Lennard-Jones tried a small number of choices for the exponent m, with 12 fitting best (to data of argon). Having made these choices, the Lennard-Jones (12,6) potential has two remaining adjustable parameters (ε and σ), see figure 1. These parameters have physical meaning[20], but this meaning is not independent from the parameterization schema, i.e., they cannot be measured or otherwise assessed independently of the entire parameterization schema. The parameter values are fitted to data for argon[21], more precisely: they are chosen according to the overall fit to training data.[22] The ansatz only becomes a model of argon after fitting its parameters to data for argon. Consequently, the resulting numbers will depend on the choice of that data and the way the fit is carried out.

---

[19] Rowlinson (2002) provides a wealth of original literature on the development.

[20] The parameter ε corresponds to the "depth of the potential well" (see figure 1) and σ to the particle size.

[21] Commonly, experimental data are used for this purpose. An alternative is using quantum chemical data for the interaction energies of argon atoms to determine the numbers for these parameters. See, among others, Jäger et al. 2009.

[22] A useful differentiation is between training data and test data. The former are used for adjusting the parameters, the latter then determine how good the predictions of the (adjusted) model are.

Figure 1. Graph of the Lennard-Jones potential function: Intermolecular potential energy as a function of the distance of a pair of particles. The graph shows the“ potential well”, i.e., a favored distance between two particles where attracting and repelling forces are in balance. As the particles also have kinetic energy and are not locked in this position, they move continuously, which is known as Brownian motion.

In MD, like generally in MMS, potentials are used like building blocks. There are different model building blocks that are put together to create a model of a complex structure. Lennard-Jones put together just two blocks to create his famous model, one for repulsion and one for dispersion. Add a dipole, and you will get another, more complex model, known as the Stockmayer potential, which has three parameters (σ, ε and the dipole moment). It is also common to combine several Lennard-Jones sites to describe chain-like molecules. However, for modeling more complex molecules, other types of interactions may become important. In general, one distinguishes between intermolecular interactions (between different molecules) and intramolecular interactions between the atoms inside a molecule, e.g., different types of vibration such as stretching, bending, or torsion. All these interactions are usually described by their potential energy, i.e., described by a potential.

$$
\begin{aligned}
V = \quad & V_{\text{intra}} + V_{\text{inter}} \qquad (2)\\
& V_{\text{intra}} = V_{\text{stretching}} + V_{\text{bending}} + V_{\text{torsion}} + \dots \\
& V_{\text{inter}} = V_{\text{repulsion}} + V_{\text{attraction}} \\
& \qquad V_{\text{attraction}} = V_{\text{vanderWaals}} + V_{\text{polar}} + \dots
\end{aligned}
$$

The different contributions are then simply summed up and build the *force field* that expresses the total potential energy of the (model) system.

Once the force field is defined, the forces acting on the particles and the resulting motion can be simulated. Much like in Laplace’s vision, except that the mathematical tools for solving the equations changed drastically. A large number of equations are numerically solved for the next (little) time step and then the procedure is repeated. Conceptually, this is a straightforward extension of the argon case. The main task of the simulation is to generate a sufficient number of representative configurations of the system to enable a meaningful determination of average properties. Such properties then can be compared to measurable macroscopic properties. In this way, one can achieve practically useful predictions (see, e.g., Eckl et al. 2008).

3.2 A round robin study

Overall, MD is a machinery for prediction. The target properties can be on the atomistic level, or on the macroscopic level. For the latter, statistical mechanics is used to compute macroscopic properties from atomistic configurations. In principle, one can determine any (thermodynamical) property of bulk material, or properties at the interface between different phases of materials. Additionally, one can model and predict nanoscale processes, like heat and mass transfer or nucleation, i.e., processes where quantum effects do not play a major role. This extremely wide applicability is driving the uptake of MD in various scientific and engineering disciplines as the prediction generator of choice. The primer to MD in section 3.1 pointed out that MD works with severe simplifications and also relies on parameterizations, including the adjustment of parameters. An obvious philosophical question is the extent to which MD models adequately represent their target domains.[23]

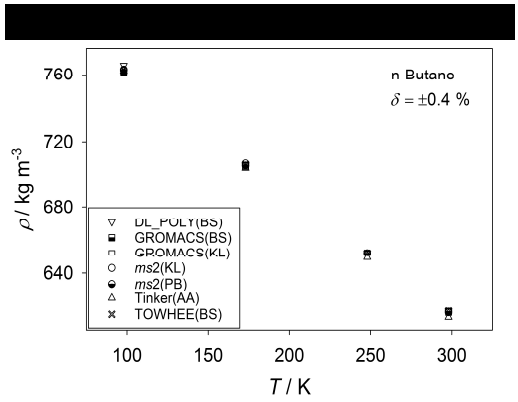
This paper does not address issues of representation and adequacy. We focus on the mathematical and computational part, or better: on the modeling process that leads from the mathematically formulated theoretical model (like equation (1) in the case of argon) to the simulated properties of a substance. According to a common view, the mathematical model determines the simulation outcome (in the limits of approximation and statistics). Consequently, the simulation is fully reproducible (again, in the limits of approximation and statistics).

We question this standard view. If the mathematical model does not determine the outcome of the simulation, reproducibility becomes an empirical question.[24] The question is relevant but tricky to answer. It is relevant, because the analysis promises new insights into computational modeling, especially into how epistemological and social aspects are intertwined. The question is a tricky one, because sufficiently documented examples of practical cases whose analysis contributes to answer the question are rare.

However, we are able to report about a scientific (engineering thermodynamics) case study in which the authors were involved (Schappals et al. 2017). It was a round robin study (different groups from different laboratories report to a central coordinating instance). The guiding question of this study was to what extent simulation experiments provide reproducible results

[23] About two decades ago, philosophical assessments of the then current nanotechnology hype discussed such questions intensely, see for instance Baird et al. (2004).

[24] Very relevant literature in philosophy of simulation will be addressed later.



when the *same* theoretical models are simulated at *different* locations by *different* groups using *different* implementations on *different* computers.

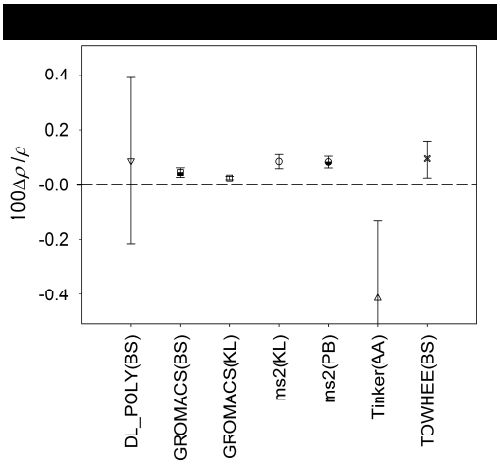
The mathematical model was clearly specified for all participating groups. One task was to compute the density (a simple property) of n-Butane (a standard substance) at various given temperatures and pressures. Five experienced simulation groups took part (four universities and one Fraunhofer institute).[25]

In the round robin study, the participating groups received the task from a central instance to where they reported back their results (hence the name round robin), however, without communicating among each other. In the first round, the simulated densities varied so widely that they were clearly worthless as predictions. It turned out that these large deviations were mostly (albeit not always) due to simple mistakes such as mistyping values. On the one hand, this sort of error is independent of the modeling process. On the other hand, this sort of errors might occur regularly in science when users handle complex codes in the error-prone way that is so characteristic of human beings – and they work on predictions in fields where it is hard to test the results independently. After some rounds with short qualitative feedback of the central instance, results were obtained that looked much better, see Figure 2.

Figure 2. Different symbols correspond to different groups that usually worked with different codes. Results are so close that symbols overlap.

However, a closer inspection of the final results gives a less clear picture. This is illustrated, like under a magnifying glass, in Figures 3+4. Here, for two temperatures, the relative

[25] Each group used the same three very common tool boxes to build the concrete force fields: OPLS (Jorgensen et al. 1984), TraPPE (Martin et al. 1998), and OPLSAmber (with stretching vibration, Jorgensen et al. (1984), Weiner et al. (1984).



deviations between the results from the different groups are displayed (the reference was the average value of all results). In such an analysis, one cannot expect perfect agreement (on the contrary: a perfect agreement could even indicate fraud by copying results of others). However, one would expect an agreement of the results within their error bars.

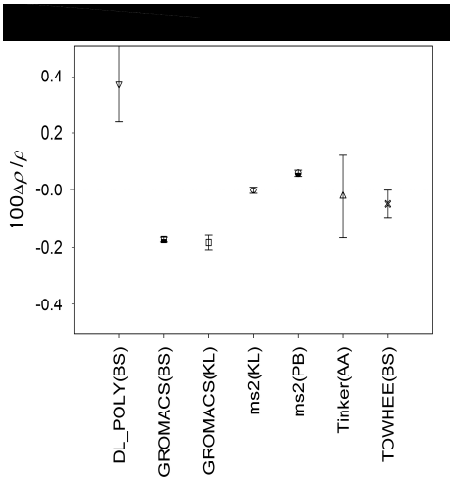


Figure 3. Results from different groups and codes for temperature 92 Kelvin, including statistical uncertainties.

Figure 4. Results from different groups and codes for temperature 248 Kelvin, including statistical uncertainties.

The results of the different groups come with statistical error bars. One can observe that the deviations between the results of the groups are larger than the statistical uncertainty of the

predictions. In other words, there are systematic differences. The groups did *not* fully reproduce (in the limits of statistics) the results of the others – even after the considerable effort of the iterations in the round robin study.

3.3 Analyzing the finding

In a nutshell, we argue that the transformation process from the well-defined mathematical model to the result of a simulation comprises many steps. This section points out that technological, epistemological, and social aspects are interwoven in these steps. As a consequence, the executable object on a given computer that produces the simulation outcome[26] is merely vaguely defined and partially opaque to the users.[27] Yes, any two groups from the round robin study started from the same relatively simple mathematical model. Nevertheless, simulating this model involves a series of mutually interacting steps and each group took their own pathway. Given that these steps matter for the outcome, two simulations by different groups will in general produce different results.

This analysis ties in with a strong direction in the philosophy of modeling and simulation. Almost from the beginning, the role of models was seen as a main factor when determining what characterizes simulations.[28] We agree with this line of thinking. Yes, models act as "autonomous agents" in Morrison's apt phrase (1999). In our simple example, the model consists of a mathematical equation (1) and the parameters. However, this does not determine

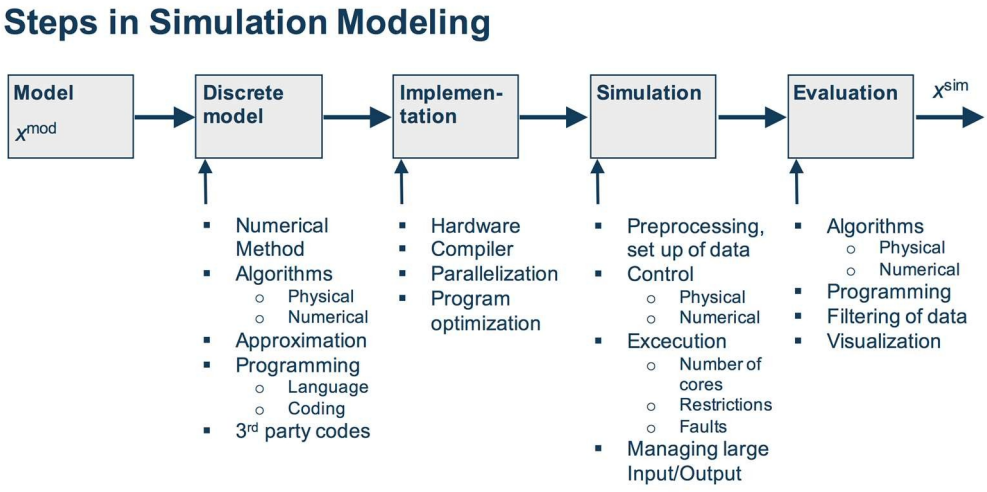
---

[26] If one calls this object the simulation model, one should keep in mind that there now are two types of model in play. The mathematical model (allegedly) represents some facet of the world, while the simulation model represents the mathematical model.

[27] The difference between user and developer is not clear-cut. Those who use simulations to make predictions, even if they rely on existing software packages, often add some developing work, too. At the same time, even a developer of some part of the software is normally also a user of other parts.

[28] Here is a small selection, spanning 20 years: Winsberg observed that simulation is "sanctioning models" (1999), Humphreys (2004) detailed a schema for computational modeling, including computational templates, Morgan (2003) and Keller (2003) argue about model-based experiments, Tal (2013) and Morrison (2015) examined how measurement interacts with (computer) models, and Lenhard (2019) analyzes simulation models as a special type of mathematical models.

the result, because a host of specifications of the steps that need to be taken to retrieve the simulation results from the mathematical model is missing, defying the apparently simple mathematical form of the model. Figure 5 gives a preliminary schema of the modeling steps



involved starting from a given theoretical model x^{mod} up to the output x^{sim} of the simulation.

Figure 5. Steps in simulation modeling. It is a long way from the mathematical model (left) to the actual simulation result (right).

There are no claims about the completeness of the schema. One goal a schema like that in figure 5 serves is to question the thesis that the mathematical model determines the simulation outcome and that, therefore, the latter is reproducible. The simulations of the groups in the round robin study of Schappals et al. differed in basically all points mentioned in the vertical columns of Figure 5. The question then is whether these differences matter. The round robin study suggests: yes.

Here is not the place to discuss all factors. A few remarks might suffice.

(1) In MD, researchers model the behavior of molecules by a large number of particles that interact according to certain rules (force fields). How many particles should they model? Are 1.000, 10.000, 100.000, or 1 million particles sufficient?[29] Obviously, this number has to be

[29] Those are typical numbers used in different types of MD simulations. All of them are very low compared to the Avogadro number 6.022 x 10^{23}.

assigned before the algorithm can start. It is an assumption that cannot be justified based on the mathematical model. It is a pragmatic assumption, depending also on technological and social factors.[30] Not every part of a simulation (in the expanded picture) can be theoretically motivated.

(2) As a rule, deviation between theoretical and simulated behavior, i.e. between x^{mod} and x^{sim}, is considered unwanted. However, normally x^{sim} is the only result that is accessible, because the behavior of x^{mod} can often not be determined without carrying out the simulation – or is only known for special cases. This initially motivates the use of simulation – and also shows how important the reproducibility of x^{sim} is, or how painful the lack of it is. Typically, simulation parameters of various sorts take on a crucial role.[31] Basically, all entries in figure 5 come with one or several simulation parameters, for instance, governing how a real number is represented in a numerical algorithm. Ideally, the simulation parameters should not influence the outcome. However, in practice this is hard (or impossible) to achieve. Importantly, the overall behavior is influenced by all simulation parameters jointly, i.e., is conditional on the entire parameterization. Any differences in parameter settings at one spot will lead to, or might be compensated by, differences in other spots. Thus, how one parameter is related to model behavior is depending on the other parameters in a highly complex way, effectively turning the influence of parameterization into a holistic function (see Hasse and Lenhard 2017, Lenhard and Hasse 2023 for an account of adjustable parameters that foster the argument).

(3) The third remark exemplifies the mutual influence of epistemic, technological, and social factors. In the round robin study, commercial software with proprietary code did not perform well. Initially, commercial codes were included in the study, and yielded poor results in the

---

[30] In general, a user will try to specify this number in such a way that a further increase will not alter the results significantly. However, other issues come into play: large particle numbers lead to lower statistical errors, but they also increase the computational effort. Hence compromises need to be made, and these compromises will depend on different factors such as the available computer, the maximal time for the simulation run that is still acceptable - or even the CO_2 emission of the simulation.

[31] These simulation parameters should be distinguished from the model parameters. In our example above, σ and ε are the (adjustable) model parameters. However, the round robin study assumed the theoretical model (including these parameters) as fixed and given.

first round. It turned out that debugging the runs was only partially possible by the users, as often input from the company's experts, who have access to the code and details of the realization of the simulation, would have been required. This input could not be obtained in reasonable time. Furthermore, such support is generally expensive. Even having access to the source code, however, may not lead to a quick solution. One would first have to become acquainted with important parts of the code. Typical MD codes have more than 100,000 lines, corresponding to about 10 books, with 250 pages each. People who have ever tried to understand the logic behind code written by someone else, are immediately aware how big the problem is. One indication is that manuals of common software packages have often a 4-digit number of pages and, still, there are few manuals that contain all relevant information. Software, including issues like maintenance and institutional access, has been identified as a crucial problem by practitioners, even as a potential "nightmare" (Miller 2006). Moreover, software has also been identified as a widely neglected topic for philosophical, historical, and sociological research (Wieber and Hocquet 2020, Hocquet and Wieber 2021, Lenhard 2014).

In conclusion, our argument is this. In order to make quantitative predictions of relevant quantities, one needs complex models, which have to be evaluated with complex procedures. These in turn can be considered "autonomous" steps, requiring the specification of parameters (or degrees of freedom). There are so many of them, and they interact, so that no one can claim to have an overview of everything. Thus, striving to improve the predictive power of the mathematical model requires to enter the thicket of simulation. The round robin study illustrated that the unavoidable differences in how different groups specify their simulation leads into problems with reproducibility. In short, these problems with reproducibility are a cost of prediction.

4 The identity crisis of models – the crumbling foundation of reproducibility

Up to here, the issue of reproducibility occurred as a problem, something science addresses during its normal business hours. And not something that signals a crisis. This final section argues that there is a crisis on the level of models, namely an identity crisis. In normal parlance, it is the model that achieves a prediction (through simulation methods). Or rather, with a philosophical touch, it is the theoretical content of the model that allows to make predictions. However, triggered by the expanded picture of the simulation process (figure 5—and constructing the initial mathematical model is not even part of the picture), one might ask: What is it that actually achieves the predictions? The mathematical model is but one component. Breeding toward prediction creates a lot of slightly different types of simulations.

All these types achieve prediction with the help of mathematical tools. But being mathematical does not imply being general and being reproducible.

It is crucial to differentiate this point of view. The theoretical model is formulated mathematically. This is an entity of great generality and apparently also of perfect reproducibility. However, getting out predictions from this entity requires a host of further steps that have a somehow (surely not completely) mathematical character, too. If one includes these steps, then problems with reproducibility are imminent, as section 3 has shown.

It seems fair to attribute the predictive success to the simulation, i.e. the overall outcome after all steps in figure 5 leading to x^{sim}. In a way, the entirety of these steps is a necessary extension of the mathematical model. Let us call the entire chain depicted in Figure 5 the simulation model. But the identity of this simulation model is not entirely determined through theoretical components. Rather, it comprises parameterizations, software codes, etc. Moreover, these components do not add up in a modular way where their function remains discernible and separable. Instead, it is a convoluted process that connects the mathematical model with the simulation result. For instance, assigning one simulation parameter value might depend on prior adjustments of how the code numerically handles discretization. And vice versa. Holism is an alternative notion that captures the situation.

An additional, though related, aspect is opacity. The notion of epistemic opacity plays a prominent role in the philosophy of simulation and there exist different variants of opacity.[32] The holistic nature of the simulation process—a result of complexity like explained in remark (2) of section 3.3 above—makes it hard to attribute features of simulation behavior to features of the simulation model. In this sense, opacity is the flip side of holism. In a different sense, the users of the simulation have only incomplete access to the simulation model when certain model features are realized in a specific way in a given code (but differently in another), but parts of the code are proprietary, or the code is large and not well documented.

In sum, the simulation model is (partially) holistic and opaque. Hence it is not straightforward, and maybe not possible at all, to tell and define what this model is. As the very brief analysis in section 3.3 indicated, this identity crisis comes from combined technological, epistemological, and social reasons.

Is the identity crisis really a crisis? Well, we do not insist on a strong claim here. At least, the fuzzy and complex identity of the simulation model causes problems with reproducibility.

---

[32] According to Humphreys (2009), demarcates simulations philosophically.

The entire argumentation ran along the case of MD, an instance of "hard" science with strong theoretical fundament, formal methods, and precision measurements. Certainly, the findings generalize to a wider field of computational approaches—or should generalize if our claims hold water. A recently extremely prominent class of examples is (deep) machine learning. It is an extreme case of parameterization. State of the art deep neural networks use billions of adjustable parameters. These networks can produce astoundingly good predictions. At the same time, they exhibit an irritatingly small robustness. For example, their predictions might depend on the random initialization of the model before it learns from (large amounts of) training data, or from minor distortions in the training data. If so, reproducing the results is extremely difficult. Furthermore, many of these models evolve in continuous training and previous versions generally cannot be retrieved, so that a reproduction is impossible for technical reasons alone. These networks are bred for one particular purpose—prediction. While they excel there, issues like robustness and reproducibility are compromised. In this way, the curious situation of P -> non-R arises. Thus, this sort of compromise should definitely be counted among the perils of prediction.

References

Abraham, M. et al. 2019. Sharing Data from Molecular Simulations. *Journal for Chemical Information and Modeling*, 59, 4093−4099.

Atmanspacher, A. and Maasen, S. (Eds.) 2016. *Reproducibility. Principles, Problems, Practices, and Prospect*. Hoboken, N.J.: Wiley.

Baird, Davis, Nordmann, Alfred, and Schummer, Joachim. (Eds.) 2004. *Discovering the nanoscale*

Battimelli, Giovanni; Giovanni Ciccotti, Pietro Greco (Eds.) (2020). *Computer Meets Theoretical Physics. The New Frontier of Molecular Simulation*. Cham: Springer Nature.

Brennecke, J.F. et al. 2019. Highlighting 10 Years of NIST Cooperation and Service to the Thermophysical Properties Data Community, *Journal of Chemical & Engineering Data*, 64, 4191−4192.

Cummings, P. T., et al. 2009. Joint Statement of Editors for The Journal of Chemical Thermodynamics, Fluid Phase Equilibria, International Journal of Thermophysics, Thermochimica Acta, and Journal of Chemical Engineering Data, *Fluid Phase Equilibria* 276, 165–166.

Dijksterhuis, E. J. 1961. *The Mechanization of the World Picture*. Oxford: Clarendon.

Eckl, B., Vrabec, J., & Hasse, H. (2008). On the application of force fields for predicting a wide variety of properties: Ethylene oxide as an example. *Fluid Phase Equilibria*, 274(1-2), 16-26.

Epstein, Joshua M. and and Axtell, Robert L. 1996. *Growing Artificial Societies*. Brookings Institution Press.

Fox, Robert (1990). Laplacian Physics, in: Olby, R.C. et al., *Companion to the History of Modern Science*, London and New York: Routledge, pp. 278-294.

Gramelsberger, G. (Ed.). (2011). *From science to computational science. Studies in the history of computing and its influence on today's sciences.* Diaphanes.

Gygli, G. und J. Pleiss. 2020. Simulation Foundry: Automated and F.A.I.R. Molecular Modeling. *Journal for Chemical Information and Modeling*, 60, 1922−1927.

Hasse, H. and Lenhard, J. (2017). Boon and Bane. On the Role of Adjustable Parameters in Simulation Models, in: Lenhard, J. and Carrier, M. (eds.): *Mathematics as a Tool. Tracing New Roles of Mathematics in the Sciences.* Boston Studies in the Philosophy and History of Science 327, 93-115.

Hocquet A. & Wieber F. (2021). Epistemic Issues in Computational Reproducibility: Software as the Elephant in the Room. European Journal for Philosophy of Science, 11(2),

Horsch, M., Hasse, H. et al. 2020. Semantic Interoperability and Characterization of Data Provenance in Computational Molecular Engineering, J*ournal of Chemical & Engineering Data*, 65, 1313−1329.

Humphreys (2004). *Extending Ourselves. Computational Science, Empiricism, and Scientific Method*. New York: Oxford University Press.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169, 615–626.

Ioannidis, John P. 2005. Why most published research findings are false. PLOS Medicine 2/8, e124.

Jäger, B.; Hellmann, R.; Bich, E.; Vogel, E. Ab initio pair potential energy curve for the argon atom pair and thermophysical properties of the dilute argon gas. I. Argon-argon interatomic potential and rovibrational spectra. Mol. Phys. 2009, 107, 2181−2188.

Johnson, Ann and Lenhard, J. (2024). Cultures of Prediction: How Engineering and Science Evolve with Mathematical Tools. Cambridge, MA: The MIT Press, to appear.

Jorgensen, W. L. et al., J. Am. Chem. Soc. 106 (1984) 6638-6646

Kampouridis, Stelios.'On the threshold of a new era': The quantitative predictive turn in Quantum Chemistry". This volume

Kampouridis, Stelios (2022). Bytes as Test Tubes: The Emergence of Computational Quantum Chemistry, PhD dissertation, University of Athens.

Keller, Evelyn Fox, Models, Simulation, and 'Computer Experiments' , in: Radder, Hans (ed.), The Philosophy of Scientific Experimentation, Pittsburgh: University of Pittsburgh Press, 2003.

Kofke, D. A. et al. 2016. Editorial: Molecular Modeling and Simulation in JCED, Journal of Chemical & Engineering Data, 61, 1−2.

Kuhn, Thomas S. (1970). The Structure of Scientific Revolutions. Chicago: University of Chicago Press.

Lejaeghere, K. et al. 2020. The uncertainty pyramid for electronic-structure methods , in Uncertainty Quantification in Multiscale Materials Modeling, Eds. D.L. McDowell and Y. Wang, Elsevier.

Lejaeghere, K. et al. 2016. Reproducibility in density functional theory calculations of solids. Science, 351, 6280, aad3000-1 – aad3000-6.

Lenhard, J. (2019). Calculated Surprises. A Philosophy of Computer Simulation. New York: Oxford University Press.

Lenhard, J. (2014). Disciplines, Models, and Computers: The Path To Computational Quantum Chemistry, *Studies in History and Philosophy of Science* Part A, 48, 89-96.

Lenhard, J. and H. Hasse (2023). „Traveling with TARDIS. Parameterization and transferability in molecular modeling and simulation." Synthese 201(129). https://doi.org/10.1007/s11229-023-04116-3

Lenhard, Stephan and Hasse in preparation

Lenhard, J. and Winsberg, E. (2010). Holism, Entrenchment, and the Future of Climate Model Pluralism, Studies in History and Philosophy of Modern Physics, 41, pp. 253-262.

Levi, Primo 1975. *The Periodic Table*

Martin, M. G. et al., J. Phys. Chem. B 102 (1998) 2569-2577

Miller, G. , Science 314 (2006) 1856–1857. A Scientist's Nightmare. Software problem leads to five retractions.

Morgan, Mary S., Experiments Without Material Intervention. Model Experiments, Virtual Experiments, and Virtually Experiments, in: Hans Radder (Hg.), The Philosophy of Scientific Experimentation, Pittsburgh: University of Pittsburgh Press, 2003, 216-235.

Morgan, M. and Morrison, M. (1999). Models as Mediators. Perspectives on Natural and Social Science. Cambridge: Cambridge University Press.

Morrison (2015). Reconstructing Reality. Models, Mathematics, and Simulations. New York: Oxford University Press.

Morrison, Margaret, Models as Autonomous Agents, in: Morgan, Mary und Margaret Morrison (Hg.), Models as Mediators, Cambridge: Cambridge University Press, 1999, 38-65.

Müser, M.H., Dapp, W.B., Bugnicourt, R. et al. Meeting the Contact-Mechanics Challenge. Tribol Lett 65, 118 (2017). https://doi.org/10.1007/s11249-017-0900-2

NASEM. US-National Academies of Sciences, Engineering, and Medicine. 2019. Reproducibility and Replicability in Science. Washington, DC: The National Academies Press.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. Science 349, 10.1126/science.aac4716.

Pashler H, Harris CR. 2012. "Is the Replicability Crisis Overblown? Three Arguments Examined". Perspectives on Psychological Science. 7 (6): 531–536. doi:10.1177/1745691612463401

Patton, Lydia (2023). Fishbones, Wheels, Eyes, and Butterflies: Heuristic Structural Reasoning in the Search for Solutions to the Navier-Stokes Equations. In: L. Patton and E. Curiel (eds.), Working Toward Solutions in Fluid Dynamics, Springer Briefs in History of Science and Technology. Cham: Springer.

Plesser, H. 2018. "Reproducibility vs. Replicability: A Brief History of a Confused Terminology" Frontiers in Neuroinformatics, 18 January 2018, Volume 11 - 2017. https://doi.org/10.3389/fninf.2017.00076.

Rowlinson, J.S. (2002). Cohesion. Scientific History of Intermolecular Forces. Cambridge University Press.

Schappals, M.; Mecklenfeld, A.; Kröger, L.; Botan, V.; Köster, A.; Stephan, S.; García, E.; Rutkai, G.; Raabe, G.; Klein, P.; Leonhard, K.; Glass, C.; Lenhard, J.; Vrabec, J.; Hasse, H. (2017). Round Robin Study: Molecular Simulation of Thermodynamic Properties from Models with Internal Degrees of Freedom. Journal of Chemical Theory and Computation 13, 4270-4280. Published online: DOI: 10.1021/acs.jctc.7b00489

Shiffrin R. & S. Chandramouli. 2016. Model Selection, Data Distributions and Reproducibility. Reproducibility: Principles, Problems, Practices, and Prospects (hrsg. von A. Atmanspacher & S. Maasen). Hoboken, N.J.: Wiley, 115–140.

Tal, E. (2013). Old and New Problems in Philosophy of Measurement. *Philosophy Compass*, 8/12, 1159–1173.

Thompson, M. W. et al. 2020. Towards molecular simulations that are transparent, reproducible, usable by others, and extensible (TRUE), *Molecular Physics*, 118, 9-10.

Weiner, S. J. et al., J. Am. Chem. Soc. 106.3 (1984) 765-784

Weisberg, Michael 2013. Simulation and Similarity. Using Models to Understand the World. New York: Oxford University Press.

Wieber, Frédéric and Alexandre Hocquet (2020). Models, Parameterizations, and Software: Epistemic Opacity in Computational Science. Perspectives on Science, 28 (5), pp. 610-629.

Winsberg, Eric 2010. *Science in the Age of Computer Simulation*. Chicago, IL: The University of Chicago Press.

Winsberg, Eric 1999. Sanctioning Models: The Epistemology of Simulation. Science in Context, 12(2), 275–292.

Wise, M. Norton 2004. Growing Explanations. Historical Perspectives on Recent Science. Durham and London: Duke University Press.