Long-Timescale Simulations: Challenges, Pitfalls, Best Practices, for Development and Applications

Graeme Henkelman, Hannes Jónsson, Tony Lelièvre, Normand Mousseau, and Arthur F. Voter

Abstract

In this chapter, we examine the practice of developing, implementing, and applying long-timescale simulation methods. In contrast to standard molecular dynamics, the performance, and sometimes the accuracy, of long-timescale atomistic methods can vary greatly from one application to another. Therefore, for the practitioners, it is particularly important to understand the strengths and weaknesses of the methods, in order to assess their respective potential for specific problems, as well as maximize their efficiency. For the method developer, clearly assessing the challenges faced by current methods as well as the areas of opportunities for future development is paramount.

In the following, we present the opinion of five leaders in the field regarding best practices, challenges, and pitfalls in the use and development of such

G. Henkelman
Department of Chemistry and Biochemistry, University of Texas at Austin, Austin, TX, USA
e-mail: henkelman@utexas.edu

H. Jónsson
Faculty of Physical Sciences, University of Iceland, Reykjavík, Iceland
e-mail: hj@hi.is

T. Lelièvre
Cermics, Ecole des Ponts ParisTech, Champs-sur-Marne, France
e-mail: tony.lelievre@enpc.fr

N. Mousseau
Department of Physics, University de Montréal, Montréal, Canada
e-mail: nm@normandmousseau.com

A. F. Voter (✉)
Theoretical Division T-1, Los Alamos National Laboratory, Los Alamos, NM, USA
e-mail: afv@lanl.gov

methods. Their answers both provide a roadmap to how best to approach the field and deliver insight into areas that need addressing in the future.

Question 1: In your experience, what are the biggest challenges associated with extending the timescale of atomistic simulations? What pitfalls does one need to be aware of?

AFV: Biggest challenges: Low barriers, low barriers, low barriers! In general, the lowest barriers in a system may be substantially lower than the relevant higher barriers, those barriers one needs to surmount to reach the timescale of interest. In this case, although there may be good acceleration compared to direct molecular dynamics, the gap between the fast rates (low barriers) and slow rates (high barriers) might prevent observation of the desired very-long-time dynamics, because a huge number of the lower-barrier events need to take place before a high-barrier event occurs. Especially frustrating is when the lowest barriers are actually so low, relative to $k_B T$, that there is very little absolute acceleration.

The very worst case is "persistent low barriers," in which the system keeps visiting new states, and each of these new states has low-barrier pathways that inhibit acceleration. For a state that has been visited previously (especially if it was visited many times previously), accumulated information can be used to improve the performance on the revisit. This cannot be done very effectively on the first visit, however, so that for this type of system the acceleration will be very low.

For systems with a serious low-barrier problem, there is hope for being able to combine many states that are connected by low barriers into a single superbasin state. The quasi-static distribution (QSD) formalization of parallel-replica dynamics offers a framework for this kind of generalization. In some cases, we have been able improve the boost substantially by using our understanding of the system to apply such an approach. So far, however, it has proved challenging to develop an automated approach to this problem, one that would work, e.g., for proteins.

NM: I totally agree with this problem. At one point, however, the low-barrier problem mixes with the configuration entropy problem. In the case of proteins, for example, many barriers have a significant entropic contribution, meaning that the transition state cannot be well described with a single point on the energy landscape.

In materials, where the problem is simpler, we have to consider two types of low barriers: those that do not evolve the system, which are generally referred to as flickers, and those that are an intrinsic part of the evolution. Distinguishing between those requires a deep knowledge of the problem. Moreover, it is not always possible to draw a clear line between the two.

For kinetic Monte Carlo (KMC), the main problem is that every time one leaves a basin, it becomes necessary to reconstruct it, which means a lot of effort for a relatively short time step. When barriers remain energy-activated, it is nevertheless possible to see a solution where basins, incorporating multiple states separated by low barriers, are generated on a local basis and merged if needed. As long as those states are distinct, KMC is efficient. When this is not the case, however, one can get overloaded by the cost of rebuilding the basin.

AFV: Unfortunately, pitfalls also abound. For example, the shape of the bias potential in hyperdynamics is a very subtle problem. In hyperdynamics, one must take care to design a bias potential that is zero at all dividing surfaces, so that it does not "block" any pathways (i.e., slow them down). A problem is that there is typically not any obvious signature when a pathway is being blocked. In principle, one can check for a nonzero bias at a dividing surface for an escape mechanism that has been observed, but for a mechanism that has not been observed because the rate was slowed down too much, no such check can be performed – then the dynamics are just wrong.

Another pitfall is the high dimensionality of typical problems in materials. Our intuition can be easily misled by drawings in one or two dimensions when the real system is 3N dimensional. An example is the flat bias potential of Steiner, Genilloud, and Wilkins (Phys. Rev. 1998), in which the potential energy is replaced with a flat potential for energies lower than some threshold, with this threshold chosen to be lower than the energy of the lowest saddle point bounding the state (picture a frozen lake in the mountains). In one dimension, or a few dimensions, this simple form of bias potential can indeed give good acceleration, and it is natural to think that this characteristic would persist to high-dimensional systems. However, it does not. Although it is a valid bias, because the bias is guaranteed to be turned off whenever the potential energy is higher than the lowest barrier, it is no longer an effective bias because the typical potential energy in a system with N atoms is roughly $3Nk_BT/2$, which for large N will be much higher than the threshold energy, so that the instantaneous boost is rarely turned on. This is difficult or impossible to draw in a one-dimensional diagram.

Another example of where our intuition can fail us in higher dimension is for a bias potential that is constructed from space-filling objects in a low-dimensional collective-variable space, as in an approach based on metadynamics, for example. Then, although the acceleration can remain large as N increases, the bias form itself may no longer be valid for hyperdynamics. This is because any reaction coordinate that is orthogonal to the collective variable space will be blocked. Thus, extreme care must be exercised.

NM: This problem is not limited to hyperdynamics. Methods such as ART nouveau and the dimer evolve the system through this high-dimension landscape to find saddle points. Yet, this search does not always work and, often, it is clear that the failure to find a saddle point is related to the structure of the energy landscape, a structure that is almost impossible to figure out given the high-dimensionality of the problem. Similarly, without a detailed knowledge of the energy landscape, it is not possible to ensure that these methods can find, even in principle, all connected saddle points.

HJ: In my opinion, the most basic pitfall is to assume some transition mechanism(s), some reaction coordinate, and base the time acceleration scheme on that assumption. This can lead to incorrect time evolution of the system. There is a myriad of cases where the mechanism for atomic scale transitions turned out to be entirely different from what one might guess a priori. The simulation should show what the relevant transition mechanism is, and not rely on a preconceived notion

of the mechanism. There are many schemes that are based on such preconceived notion, metadynamics being one of them that is used frequently. It is also important to realize that full free energy sampling using such an assumed reaction coordinate is not going to make up for a bad choice of the reaction coordinate. If the sampling is carried out in a subspace (e.g., hyperplane) that is not lined up with the transition state of the relevant mechanism, then no matter how much sampling is carried out, the deduced rate of the transition will be wrong and the simulated time evolution incorrect.

Question 2: What are the main limitations of current methods and what needs to be done to address them?

NM: There has been considerable development over the last two decades regarding accelerated methods. While they have opened up new regimes of physics that had been out of reach until now, significant challenges remain.

(I) Cost-effective accurate force fields. Long-time simulations in materials typically require relatively large systems, as following kinetics often implies displacement. Ab initio calculations being still limited to boxes of 1000 atoms or less, they cannot be used directly in these simulations. For lattice-based atomistic KMC methods, it is possible to construct an event catalog using small cells and density functional theory (DFT). Yet, for complex systems with significant deformations or a large number of configurations, even catalog building becomes too costly with ab initio. When dealing with elemental systems, especially pure metals, empirical potentials can offer a decent level of precision. These potentials cannot be relied upon, however, for conformations far away from the close-packed states, in the presence of many elements or when considering semiconductors, magnetic elements, etc. Over the last decade, statistical-derived potentials, using neural networks or other automatic learning methods, have shown significant promise. There is still work to be done, however, before these methods can be used regularly. Moreover, these new approaches remain very costly, limiting their application.

(II) Entropic effects. KMC methods, whether off-lattice or lattice-based, are effectively run at zero K. Most of the time, entropic contributions are included through a constant prefactor that supposes that the local environments remain relatively similar throughout the simulation. Some groups go beyond this simple approximation and evaluate the local entropic contributions for each environment using the harmonic approximation of transition state theory (TST), which supposes a temperature-independent prefactor, an approximation also made in temperature accelerated dynamics (TAD), one of the accelerated MD (AMD) approaches. Yet, this is not always valid. For example, an atom moving in a three-vacancy cluster in an FCC metal forms a tetrahedral vacancy cluster centered around an atom, a structure stable at high temperature. As it turns out, this structure is very unstable at low temperature, although it exists. In Ni, this state is at an energy 0.4 eV above ground state, separated from it by a barrier of only 0.08 eV. Its stabilization at high temperature is clearly due to an increase in thermal vibrations that prevent the atom from moving back into its original

position, a feature that might not be captured by the harmonic approximation (S. Mahmoud, et al., Acta Materialia (in press)). It is therefore likely necessary to go beyond the harmonic approximation with thermodynamic integration methods that can be automated.

(III) Going beyond TST. Most accelerated methods rely on TST at some level. There is a need to work on this theory and see how one can expand it to complex systems without having to use the heavy tools developed in systems dominated by entropy, such as transition path sampling.

(IV) Handling flickers and energy barrier distributions efficiently. In spite of considerable advances over the last years, handling systems with a wide distribution of relevant barriers remains a challenge. Yet, following the evolution of a grain boundary, the formation of a nanostructure, or the aging of a multicomponent glass, involves working with continuous energy distributions.

(V) A more efficient cataloging. Recognizing local environments, whether using KMC or MD-based methods is an essential part of recycling previous efforts. Over the years, a number of approaches have been proposed – geometrical, lattice-based, and topological. Yet, all of them suffer from some limitations that decrease the recycling. There is a need to carefully study this aspect and identify methods that are flexible, can be applied to wide range of environments, provide useful comparison, and facilitate the reconstruction.

HJ: Current methods are mainly based on the identification of local minima on the potential energy surface (with the exception of parallel replica dynamics, though even there, in practice, states are often defined in terms of local energy minima). There is a large set of problems, where entropy plays a central role and energy barriers are small and numerous. The definition of states needs to be more in terms of subspaces where the system spends enough time to locally equilibrate and for which the time evolution can be described as a Markov chain. Full TST, as opposed to the harmonic approximation to TST, coupled with the variational principle can, I believe, deal with these kinds of systems, the problem is coming up with an efficient algorithm for representing and optimizing the dividing surface that defines a state. Very little work has been done on this so far. Is it timely now to give it a try?

TL: I think that in some sense, AMD methods can be seen as a way to go beyond the TST, especially parallel replica dynamics or parallel trajectory splicing. Indeed, the decorrelation step is a way to check if first order kinetics can be applied to model the exit event, and then, these algorithms do not require knowledge of the exit rates. However, the efficiency in parallel replica is limited by the decorrelation time within the state. Parallel trajectory splicing is a nice idea to overcome this problem, but we need to think more about algorithms that are able to exploit massively parallel machines.

AFV: Normand, Tony, and Hannes make good points. As discussed under question 1, in my mind the main limitation of the methods is how hard it is to get good acceleration for systems with persistent low barriers. Although some progress has been made on this front by various groups, I have been surprised by how difficult this challenge has remained. Ever greater cleverness will be required.

Amplifying Normand's point on the difficulty in using electronic structure forces, although in principle AMD methods can be used as easily with an expensive first-principles force call as it can with an empirical potential, in practice this may be far from the truth. For electronic structure forces, it may take so much computer time to advance the dynamics that the system will still be in its initial state when the computer budget runs out. In this situation, while formally the boost factor may be high, if the system has not yet jumped to a new state, the effective boost factor is zero.

Question 3: Are there specific issues or challenges associated with applying these methods as opposed to their development? What could be done to further their widespread adoption?

GH: Echoing Normand's point above, perhaps the most important challenge associated with a more widespread application of accelerated methods is that they cannot, at present, be used with standard DFT. There is an enormous community of scientists using DFT to model dynamics in chemical and material science applications, but we have not provided accelerated methods which are sufficiently efficient to be used routinely with energies and forces from DFT. While current methods can be based upon empirical potentials, for example as implemented in parallel codes such as LAMMPS, there are a vanishingly small number of applications for which we have accurate potentials as compared to those which can be described by DFT.

Some of the problem, I believe, can be attributed to the electronic structure community, which puts a greater emphasis on the accuracy of each energy calculation rather than an accurate sampling of configuration space. Recent developments of surrogate models, including those based upon machine learning, have the potential to change this bias if sampling could be done at a fraction of the cost of DFT. Additionally, there is a potential application for more approximate methods which can be used directly with DFT. Regardless of which direction the connection is made, either more approximation and efficient methods or more approximate and efficient potentials, providing tools in a form which can be used directly by the community of people running DFT calculations is key to their widespread adoption.

TL: I have the feeling that one of the major difficulties when applying AMD techniques in a general setting is the definition of good metastable states. It would be great to be able to define automatically or adaptively (as the simulation runs) good metastable states. These metastable states should be such that:

 (I) The time to leave these states is much larger than the time to reach local equilibrium (quasi-stationary distribution QSD), for a generic initial condition obtained when entering the state.
 (II) The discrepancy between the transition rates associated with these states is not too large (otherwise the algorithm spends much of the time switching over low barriers).
(III) The states give a reasonable description of the macroscopic state of the system (since the details of the dynamics within states are lost).
(IV) There is a way to estimate the convergence time to local equilibrium faithfully.

The list is rather long, but one could think of using modern statistical techniques to extract from the trajectories "good" states. The fact that these states do not need to define a partition of the state space could be used to make the construction easier. This is very much related to discussions about choosing good reaction coordinated or collective variables, which is obviously a difficult question, but the interest of AMD methods is that even if some of the states are not really well chosen, the methods can still give reliable results, e.g., with parallel replica dynamics.

AFV: I have been somewhat surprised that effective application of the AMD methods seems to require serious dedication on the part of the user. In essence, the user must become an expert; this takes time, and some users become discouraged before reaching this stage. Further automation of the methods, as discussed above, should help, and we may achieve this in the future, so that for a nonexpert user, applying an AMD code could be a "turnkey" operation.

NM: In addition to the challenges already mentioned, with which I totally agree, I would add that the current accelerated methods, based on high energy barriers with respect to $k_B T$, are restricted to solids well below melting. This is a considerable limitation as it prevents us from looking at molecules in solvent, including biomolecules, many catalytic and growth processes and a number of other fundamental questions. Overcoming this limitation is not impossible but it will require rethinking the approach and, more important, reworking TST.

HJ: Transition state theory actually works fine in cases where the free energy barrier is mainly of entropic origin. The effusion through a hole in a cavity is a nice example of how TST can give exact results even where there is no energy barrier. What is missing, however, is an efficient implementation of full TST where the dividing surface is systematically optimized (using the variational principle) to obtain the free energy barrier and thereby the mechanism(s) of the transition. But, I want to emphasize that I agree with all that has been written here above.

Question 4: In developing or applying an approach, what are the best practices you recommend?

HJ: The first rule is to have one or more test problems that are simple enough or well enough established so that the answer that should be obtained from the calculation is known. For example, when testing a method for estimating a transition rate, choose a system where the energy barrier is low enough and the temperature high enough that relatively long, but not too long, simulations of the dynamics using basic equations of motion can be used for comparison. When finding saddle points, use a test system where the energy surface can be visualized and the search path illustrated. Also, when it comes to implementation of a previously developed approach, do a calculation on a system that has been studied with the method previously and where the performance has been documented. Here, the web site OptBench.org is of great help. There, various benchmark problems have been documented and the performance of various methods reported by experts. There are too many articles in the literature where performance of a method is reported but the implementation is not optimal. It is important to compare performance reported by those who developed or are knowledgeable about the method.

The challenge in atomic scale simulations is typically the large number of degrees of freedom. While a system with only a few degrees of freedom can be valuable to test and illustrate a method, performance should not be measured with application to such systems. A method that works well for a system with only a few degrees of freedom may not work well for a realistic system with many degrees of freedom. The opposite can also be the case. It is important to document performance on challenging systems for which the method has been developed.

When documenting performance, it does not make sense to report CPU time. Computers change rapidly and such information is quickly obsolete. Identify the most computationally intensive operation and report the number of such operations needed to reach the desired results from the simulation. In most simulations of atomic scale systems, the evaluation of the energy and atomic forces is the most computationally intensive operation. A natural measure of computational effort in calculations of transition rate, identification of a reaction mechanism or a saddle point search is the number of times the energy and force needs to be evaluated.

It is also important to keep focus on the ultimate goal of the calculation, not intermediate steps. The question whether an approach is useful and how large the computational effort is should be answered by evaluating the quantity of interest.

Performance in terms of computational effort is of course not the only criterion for evaluating the appropriateness of a method. A method that is not reliable in that the answer obtained cannot be trusted is not useful even if it is fast. The results obtained using a fast method should be compared with results obtained using a slower but safer method on a range of problems similar to the application of interest.

GH: The points raised, including the recommendations to compare new methods to existing methods through benchmarks and the strategy of developing methods on model system with known results and then demonstrating how well they work in complex high dimensional systems are spot-on. Adopting these recommendations as best practices would benefit the community of method developers and the people who aim to apply the methods. An additional recommendation, which is gaining traction in the community, is that computational methods that are developed in the public domain should be made available in the form of open source code. There are many details associated with computational methods which are not easily described in publications. To make our methods transparent and our calculations reproducible, other developers and users should be able to see the algorithms at the level of the source code and reproduce published results and benchmarks directly from the code. The adoption of an open-source policy for computational material science is, in my opinion, encouraging collaboration in the field and facilitating the development of computational frameworks that are larger than the scope of a single research group, such as the Materials Project.

Finally, I think that there is an opportunity in the field to consider how computational methods compare with respect to more than one objective. Taking the example of modeling molecular dynamics over long timescales, for example, we have efficient methods based upon harmonic transition state theory (e.g., temperature accelerated dynamics and off-lattice kinetic Monte Carlo), which have the inherent limitations associated with that approximation, and other methods for which the

harmonic approximation is relaxed (e.g., parallel replica dynamics and hyperdynamics), that may have additional computational costs. What is not typically considered is the pareto-optimal set of methods which can deliver the highest accuracy for the minimum computational cost. In other words, the community of developers can establish a set of tools which can most efficiently accelerate dynamics for a specified level of accuracy. There is similar multiobjective optimization problem between the accuracy of the energy and force evaluations (e.g., empirical potentials vs. density functional theory) and the degree to which the energy landscape can be sampled. Research groups that focus on the accuracy of electronic structure calculations can neglect the potential importance of exploring the energy landscape. On the other hand, a focus on highly accurate sampling will typically put little emphasis on the accuracy of the potential landscape. A set of efficient methods for modeling dynamics or sampling potential surfaces targeting a wide range of available sample sizes would facilitate the use of methods such as accelerated dynamics for the large community of scientists modeling systems of interest with density functional theory.

AFV: Graeme and Hannes have covered this well; I will add just a few general points. As with any careful computational work (or any careful science for that matter), one should always be on the lookout for indications that something is not working correctly or not making sense. This is especially important for simulation methods that are capable of giving results for timescales that cannot be reached in any other way. For example, perhaps there is a nonlinearity that only causes significant inaccuracy at very large boost factors, which means it might not show up until the simulated timescales are beyond what can be checked with MD.

Although I think most readers of this chapter will already understand this, there is a general principle I feel is important for developing any method of this type – a method that is tied to an interatomic potential and that attempts to improve the efficiency of the simulation of a physical, material, or chemical property. When testing the method, the accuracy should be gauged by how well the method reproduces accurate simulation results for that same interatomic potential, *not* an experimental result. For long-time dynamics methods, the correct reference is direct molecular dynamics, and the benchmark systems must be chosen with some care, as MD may not be capable of directly reaching the long timescales (as Hannes pointed out).

Finally, on this issue of common codes, while I agree that standard software packages make development faster and easier than ever, one should absolutely not be afraid of developing one's own code from scratch. This can take longer, but sometimes it opens possibilities for creativity in the development that would be steered in a different direction, or inhibited, by using existing packages. Moreover, the developer typically gains a deeper understanding of the methodology by proceeding in this way. And standard codes are not totally bug- or mistake-free, so the developer may uncover such problems by having a redundant code with which to compare, thereby doing the community a favor.

NM: Previous advice is excellent and I agree with all those. Developing methods is a risky business. It is generally impossible to tell how well it will perform on realistic systems before the method is in place. For a method to be useful, it has to

be beyond what is available, either by doing faster or allowing access to new regimes of physics. This is why, I would also suggest that you make sure that the method that you are developing can go beyond what other approaches can do by applying it, as soon as possible, on nontrivial problems. So: yes, do check on simple systems that can allow you to compare results with other approaches, but make sure that also apply your algorithm to something nontrivial that demonstrate the interest of your method; too many methods are demonstrated on simple well-understood systems but have failed to produce any new significant physics.

It is also worth remembering that few computational physicists today write extensive codes. Most will use standard packages and write analysis or extension bits. If your algorithm has any degree of complexity, it is therefore essential that you write a code that is portable and useable by others, if you want your approach to gain exposure and have the impact you expect. This means that your code should be easy to read, modular, and adaptable. Be aware of this requirement from the day you start planning your software; this will decrease the time you will have to spend getting your code ready for distribution.

TL: Let me make two points along those lines. First, I would like to mention here that it is actually very difficult for nonspecialists to have access to test problems which are simple enough to have reference values, but not considered overly simple by the applied community. It would be very useful for the development of new approaches to agree on a sequence of test problems with graded difficulty, where the problem would be only a timescale problem (and not a modeling problem related to the force field for example). Second, when very long timescales are reached, weird results may be observed because of two reasons: (i) the algorithm gives incorrect estimates or (ii) the model is incorrect. Concerning the second item, notice indeed that many force fields have been parameterized and checked using only short-timescales simulations: when looking at very long trajectories, one is thus visiting unexplored territories. In such situations, it is very important to have rigorous ways to assess the quality of the numerical results in order to distinguish between the two sources of errors. This shows the importance of deriving certified error bounds for such algorithms (which is indeed sometimes a challenge!).