10
Planning a Project

This chapter has several aims. First, to help you decide which research questions you want to answer. This is an important part of planning any scientific study, but there are aspects that are peculiar to computational studies. Second, to guide you as to which method to use in answering these questions. All computational techniques have strengths and weaknesses, and when you know what questions you are asking, you will have some idea of what approach to use to answer them. Third, to provide some guidance on how to setup the project and decide what simulations are required.

We start by considering how the overall shape of the project is formed, and in particular, how to frame the research questions that are being asked. We then move to the question of resources: planning requires both knowledge of resources and a realistic approach to their use. We discuss the model of the system, including the simulation cell, and the factors affecting the choice of a simulation method or methods and then describe what types of quantity can be reasonably calculated with different methods. We finish the chapter by examining how to document a simulation.

There are two main groups of people who will find this chapter helpful: scientists entering the field, particularly students; and experimentalists working in the field seeking to understand more about modeling. For students, you will need to learn not only the techniques to use, but also the approaches which are inherent in scientific investigation. For experimentalists, it may be that you want to know which methods you should learn how to use, or to find the right kind of theorist to collaborate with. It should also help you evaluate modeling papers in your field: how far should a particular choice of modeling protocol be trusted? How reliable is a given method ?

10.1
Questions to Consider

One of the dangers with a computational tool is that there is a temptation to use it simply because it is available, without reference to the experimental background. Just because something can be calculated does not mean that it is important or

Atomistic Computer Simulations, First Edition. Veronika Brázdová and David R. Bowler

relevant. There are too many papers published which simply present the results of a simulation: computational simulations should either interpret existing results, or predict new directions for experiment, with clear reasons and parameters. This section will help you to consider the larger context when planning simulations, and to frame the questions that you are asking. We have divided it into two areas: first, the research questions which you are trying to answer; second, questions related to the simulations that you will perform.

10.1.1 Research Questions

There are four main questions that you should be asking when planning a research project:

- What data is available from experiment?
- What do we already know from other simulations?
- What do we want to know?
- What can we calculate?

The first step in planning a project is to ensure that you are familiar with what is already known about the system or the problem. This should start with the experimental background, which is essential in any planning process. As well as researching the data which have been found by experiment, you should familiarize yourself with what can be measured experimentally. This serves two purposes: first, it will make clear what data you can expect from experiment, for example, whether the technique gives information about the elements being measured; second, it will help you plan the simulations you will perform.

Once you are confident with what is known experimentally, you should investigate the previous simulation work. You should always benchmark your methods and codes against both experimental data and existing simulations. Previous simulations will show you what methods can be applied to the system, and will also give you a starting point. They should also indicate if there are particular problems with the system you are researching.

The questions that remain unanswered in the field will provide the raw material for planning your simulations. You will need to consider what can be simulated, and what you want to know. The questions may come from experiment, from simulation or both.

Simulations can serve a number of purposes in working with experimental data. First, they can confirm a postulate which has been made about the system being studied by predicting certain observables based on that postulate. For instance, given a suggested crystal structure, NMR data or vibrational frequencies can be predicted and compared to measured data. If the results are not in agreement, and both experiment and theory are confident in the accuracy of their data, then further refinement of the postulate can lead to an improved structure. The most important thing when testing experimental data is to ensure that you do not simply reproduce the experimental data: the interpretation and understanding is also vital, and lifts simulation out of the pedestrian.

Simulations can also make predictions to lead experimental measurements. Given what is known about a system, new experiments might be suggested to test a postulate generated by simulation (in a nice inversion of the previous point), or novel structures or compounds might be suggested for synthesis and characterization. When leading an experiment, it is vital to carefully consider two questions. First, what the most useful data to model would be, rather than the easiest data. Second, the restrictions and approximations in the modeling must be carefully characterized so that it is clear how reliable the suggestion is. It would be unreasonable to expect experiments to be planned on the basis of unreliable simulation data. The effort required to obtain an experimental result should also be taken into account when suggesting new investigations: is the data really needed?

Finally, simulations can interpret and unravel experiments, providing deeper insight into mechanisms which cannot be directly measured. This role is important, and encompasses some facets of both previous points. Given experimental data, by testing postulated structures or mechanisms, simulation can explain the results. By then exploring the system further, the mechanisms leading to the experimental observations can be unraveled. The bonding leading to an observed structure, or the important interactions giving particular diffusion constants, are not directly observable in experiment, and simulation can play a vital role in opening these up, which leads to better understanding and further experiments and simulation. This interaction should be on-going, with several cycles from experiment to theory and back again. Raw data should also be shared with interpretation being a joint effort, so that both experiments and theory are applied efficiently. Ideally, computational simulations and experiment will work together so that there is no sense of either being more important, though this level of collaboration takes time to build; we discuss the comparison with experimental data in detail in Chapter 18.

These questions all fall under a broader idea: what is the aim of the simulations? It is easy to start calculations without considering the overall aim, and this can very quickly lead to time-wasting. If you have a good picture of the aim of the project, and keep this in mind while working on the project, you will be more efficient. It is inevitable that new results and problems will emerge as a project unfolds, and these may well provide interesting but ultimately irrelevant side questions. With a clear understanding of the project's aim, these distractions can be managed sensibly.

10.1.2 Simulation Questions

Once you have identified the research questions, you can turn to simulation questions. The important and necessary scales are the first questions to settle. What are the timescales and lengthscales that you will need? There will need to be, naturally, a compromise between what is needed and what is practical for both these problems. Timescales are sometimes very hard to correctly match: for instance, when considering macroscopic movements (e. g., a probe moving into the system) over the timescales of a simulation, velocities are often at least two orders of magnitude too high in the simulations. The effect of this increase in velocity must be understood. The energy scales required may also play a part in your simulation, for instance, the kinetic energy of ions, or the magnitude of excitation required, and thus how large a set of unoccupied states to include.

The types of energy which must be calculated is a key question. We can divide this into total energies versus free energies, with the latter requiring a suitable sampling approach to be used. For total energies, it is important to distinguish between thermodynamic effects (where the energies of different structures are important, and equilibrium can be assumed) and kinetic effects (where barriers will be important, and the details of sample preparation are key). How rare events are sampled is often important: they must be accelerated to be observed in a simulation, and this will bias the results.

If you are comparing to or predicting experimental results, is there any averaging inherent in the experimental method? You may need to consider how to average so that your results can be compared to experiment. If there is significant sampling required, you must also consider the size of configuration space, and the length of simulation required to explore this adequately.

The environment is important to consider and will affect the choice of boundary conditions. Are there significant external effects which must be considered, for instance, solvation, the pH of a system, or the dielectric constant of the surroundings? Are there external electric fields or temperature gradients which must be imposed? The question of boundary conditions, and the effect that your assumed boundary conditions may have is a key part of planning the project, and will be influenced by the size and structure of the system.

The question of ground state versus excited state properties can be considered both in terms of the thermal occupancy of different structures, and in terms of the electronic structure of the system. The first of these has already been covered above, in considering free energies and thermodynamics versus kinetics. The second will require care, as the calculation of excited states is more complex than ground states, and generally requires more computational resources. Basis sets and accuracy need careful testing.

Accuracy is an extremely important question to consider at an early stage, though it should be distinguished from precision. The *precision* of a calculation relies on the computational parameters chosen (and is discussed in Chapter 12), whereas the *accuracy* is a property of the method, and relies on the inherent approximations made.

The accuracy of a method, particularly for calculating the quantities you have identified as being important, will determine whether it is suitable for the project. The accuracy required will also depend on the available data from experiment, and whether you are interpreting existing data or predicting new data to lead experiment. If there are different possible structures or mechanisms you are trying to distinguish, then you should consider what accuracy is required to do this unambiguously. There will certainly be times when an approximate answer which can be obtained quickly will be more valuable than an accurate answer found slowly;

you will need to examine the resources you have available when deciding how to simulate the problem. In particular, more accurate methods tend to be computationally more demanding, which limits the feasible system sizes and therefore may introduce boundary effects. Method choice and system size will be opposing requirements.

10.2 Planning Simulations

10.2.1 Making it Simple

It is tempting to try to simulate a whole process, such as a chemical reaction, protein folding or crystal growth by a dynamical simulation. However, in most cases, it is both unnecessary and impossible. Instead, think of the key configurations your system can sample, the configurations that must be included to correctly describe the events happening in the system. An analogy in the macroscopic world is, for example, Eadweard Muybridge's 1872 proof that a horse will have all four feet in the air at some point when galloping. Muybridge's groundbreaking series of still photographs taken at different stages of a horse's gallop showed one snapshot where the horse did indeed have all four feet in the air. Without the need to create a movie, indeed before the invention of motion pictures, the question was settled.

Similarly, you should avoid unnecessarily complicated simulations. For instance, it may be possible to avoid simulating the full dynamics of a system by only optimizing the starting and ending configurations and finding a transition state. Or, it may be sufficient to perform MD using a forcefield, and then make single-point calculations of snapshots with an electronic structure method. Of course, it will not be always possible, but it is worth a try. You will certainly need to begin with static simulations even if you will then proceed to do any kind of dynamics.

Once you have identified the key configurations, consider whether they are minima or transition states on the potential energy surface. How many of them do you need to simulate? Can any be omitted without affecting the correct description of the process? Which configurations can be calculated with the information you have and which of the rest depend on these? For example, you may know the structures of reactants and products of a chemical reaction from experimental data, and therefore calculate the corresponding optimized structures relatively easily, with two simple structure optimizations. A search for the transition state of the reaction, however, will depend on having the configurations of the reactants and products.

10.2.2 Planning and Adapting the Sequence of Calculations

There is not necessarily a right way or a wrong way to perform simulations, but there are certainly approaches which will help to speed up the process. There is no point in running calculations on the largest model you may have, with the highest precision from the start: doing so will almost certainly waste computational time. A sensible approach is to characterize the system using relatively small calculations at moderate precision, which will help your understanding of both the system and its behavior. Once you have this understanding of the system, then you can plan your more detailed simulations.

A preliminary investigation should seek to characterize the important scales in the system: energy scales, timescales and lengthscales. When you have some understanding of these scales, you will be able to decide on the method needed to achieve appropriate accuracy, create the model with the necessary size, and plan appropriate dynamics. You should then seek to characterize the important interactions and areas of the simulation cell. For instance, the adsorption sites for an atom or molecule on a surface, or the flexibility of a protein. These characterizations should be relatively simple, cheap simulations: you are not generating detailed data, but planning how you will generate that data.

There is a necessary balance between the broad overview and the detailed investigations. Computational tools give enormous power to explore the effects of changing numerous parameters, and generating vast amounts of data which may be of little use or interest to anyone else. Hence the need for a good understanding of the background, and the broad overview of the system: keeping this in mind when performing detailed simulations is vital. Stepping back from the everyday wrestling match with simulation codes and computers will help you to keep things in perspective: how are your results fitting into the wider picture? Are you spending too much time on something which is not important?

Your simulations should be planned carefully. When relaxing a novel structure with large forces, you do not need as precise forces as when approaching the end of a relaxation. Similarly, initial calculations can use smaller simulation cells than you would use in the final calculations, to ensure rapid characterization and convergence. It is perfectly possible to start small and reuse some data in a larger cell or a longer simulation. Thinking about these areas will save you time and result in better simulations. Similarly, tolerances do not need to be high in all simulations: adjust them according to what you are doing, and the stage of the calculation.

It is important to revisit your plans regularly. While initial simulations may have suggested one area as being important, your detailed simulations may have brought up unexpected new data. Or it may be that one set of simulations is less interesting than expected, or disproves an initial postulate. Under these circumstances, it is far better to revisit plans and adapt than to keep going along a flawed pathway.

Regular consistency and convergence checks are also important. It is very easy to set parameters wrongly, and consistency checks will catch this kind of error: it is better to catch it early. It is important to have consistency checks against experiment and previous simulation results as well as against your own simulation results. If there is a simple way to test that the change you have made in a simulation cell or a system is sensible, then you should do. Make sure that doubling the number of atoms doubles the energy, and that changing the timestep gives little change in velocities and energy conservation. If you introduce a new set of atoms, do you have the coordinates laid out properly?

There are many types of convergence. Often, energy differences and forces will converge faster than absolute energies, particularly in a variational approach. You should check that this is as expected, and test all the assumptions you have made about system size, timescales and other parameters. How does your simulation cell affect your results? While you may not be able to achieve full convergence, you must describe how the convergence will affect your results.

10.3 Being Realistic: Available Resources for the Project

An essential part of planning a project is to consider the resources available to you, and the requirements of the problem you will be investigating. We have already discussed the simulations which will be required, but these must be kept in mind as they will make specific demands on the methods required. The resources that you should consider are:

- Total time for the project
- Computing power available
- Methods and computer codes you can use
- Support for computing

These resources will affect the simulations that can be performed during your project. It may be that, given the available time, the approach you would like to take is not feasible, or that you do not have the computational resources or an available code that you would need to make the calculations in a reasonable time.

Often, the most important resource is that of the time available to you. If your project has a hard time limit (say for a Master's degree), then this will constrain the types of simulation that you can perform. Other examples of this include funding being available only for a specific time, and a collaborator or employer requiring an answer by a specific date. While this may seem rather an obvious point, it is often overlooked. You should also consider contingency planning, particularly for a time-limited project. A regular review of progress against expected achievements will enable you to address any problems with the project. If you face unexpected loss of computer time, or poor progress in the project, a change in the aims and replanning given the available time will enable you to finish the project with some success.

Computational resources are described in detail in the Appendix B. They fall roughly into two divisions: the raw CPU power available to you, which will constrain the size of the simulations you can perform, and affect the speed at which you can run simulations; and the total CPU time available. If you run calculations mainly on workstations, then the CPU time is set by the time available to the project and the number of workstations you can commandeer; on high performance computing facilities, the CPU time is almost always charged or limited, and this will form the resource.

The raw CPU power will determine how long a simulation of a given size takes to run. The use of parallelization and HPC centers can reduce this time by using more processors, though the speed up from parallel scaling is rarely perfect. HPC centers inevitably involve use of a queuing system, and you may have to wait for your job to run. Moreover, running on more processors will use more total CPU time, which is normally strictly limited. A local resource, while slower, is more under your control. You will probably want to balance the need for rapid response against the ability to model large systems.

When considering the methods required, you should characterize how well you understand a particular method, and also the code in which it is implemented. These are not necessarily the same thing. Moreover, it may be that the method you require has not been implemented in any computer code. In that case, you will also need to account for the time required to implement and debug the method, or any variation you make to the method. If the method is available, but is new to you, you will require time to learn how to use it. The state of the code is very important. While there are many codes which are stable and well-documented, there are also experimental, poorly documented codes, and every variation in between. Techniques which have been recently developed are more likely to be poorly documented and less stable.

Computational codes are often available freely within the scientific community. As these codes will normally have been created by a researcher, the level of support and documentation will vary wildly from code to code. It is a good idea to choose a code which at the very least has an active community of users, as this will often be most helpful in solving problems. Freely available codes will generally have been provided with the source code: this has both good and bad sides. The good side is that you can develop and debug the code yourself, which can be a powerful tool. The bad side is that you will probably have to compile the code yourself, which can be a significant challenge. The computational support available to you locally as well as in the community of the code will be an important factor here. There are also an increasing number of codes that are charged for, and which may or may not come with source code. These often include support as part of the cost, and this will be a factor to consider.

Computational support is an important resource to consider. Do you have support for maintaining the system on which you will be running the simulations? This might include both support for the system itself, and for compiling and optimizing the code. Local or national computational support can be extremely effective when trying to find problems with a code. At the same time, being part of an active community of users, whether within your group or across the world, is an important part of success with computational simulations: when things go wrong, which they will, it will take time and effort to fix the problems, and this will be easier and faster with a community.

10.4 Creating Models

A model is a representation of the experimental system that you are wanting to study, using a finite number of atoms and a given simulation method. The method or methods you choose will dictate the size of system that you can simulate, the properties that you can calculate and the accuracy with which they are found.

Most of the work involved in creating an appropriate model is related to finding or predicting the coordinates of the atoms you will include, and this subject is covered in detail in Chapter 11. You can often find starting coordinates from experimental input, though there will be work to do before you can use them. For instance, you may have to add hydrogen atoms to a protein structure, as X-ray crystallography rarely identifies the hydrogen positions. Experimental results may be able to tell you the symmetry of a system or the approximate location of the structure of interest, but not the detailed atomic positions. In these circumstances, you will need to work carefully with experiment to develop plausible structures which you can then test against available data.

You will also need to carefully consider the size of the computational system that you are modeling, and the effect of boundaries. A small system will, naturally, take less time to model than a large system, but may be less accurate. There may be interactions between periodic images, or unexpected confinement effects. Exploring the effect of your model size on total energies and other properties is an important part of any investigation. However, a small model may well capture most of the important science without making bad approximations, and can serve as an important tool in exploring the properties of the system.

Typically, preliminary investigations are carried out using a small model to characterize the system and its properties. These calculations can be made with gradually increasing precision, as the details start to emerge. Once you have identified the interesting science and the characteristic length and timescales of the problem, you can go back and revisit some calculations with larger simulation cells, longer times, higher precisions or any other changes needed to establish the accuracy of your results.

As an example, consider calculating the simple vacancy in bulk silicon. The smallest computational cell for silicon is a two atom fcc unit cell, while the smallest orthorhombic cell is cubic, with eight atoms. Making a vacancy in either of these would result in very small distance between defects, with unphysical effects resulting, so a larger simulation cell is necessary even for the preliminary calculations. A $2 \times 2 \times 2$ aggregate of the cubic cell gives a simulation cell with 64 atoms which is a little over 1 nm on each side, and should be large enough for preliminary calculations. However, testing the properties you have calculated in larger cells is important: you should certainly look at a $3 \times 3 \times 3$ aggregate (with 216 atoms) and move to larger cells if necessary.

When considering a less homogeneous system, for instance, a protein, it may be less clear how the model should be created. Consider an enzyme with an active site: the atoms which are functional in the active site must be included as well as the neighboring amino acids. There are some investigations where a model this large may not be feasible, and substituting a small molecule with similar properties for a large one may be required. You can define sets of atoms which contribute more or less to the properties of the system you are studying, and investigate how important these sets are to the overall results. If cutting bonds, say the amide bond between amino acids, a chemically appropriate termination is required.

10.5 Choosing a Method

Probably the first question to ask about the method to use is whether the electrons need to be included explicitly. This differentiates forcefield methods from electronic structure methods, and broadly sets the level of computational effort required. It is not an easy question to answer: it depends on how important electronic effects such as bond breaking and making or charge transfer will be in the problem, and on the scale of system you are simulating.

If you do not require electronic structure, then you need to examine the forcefield that you will use. All forcefields are fitted to specific data, and this will set the range of validity of the method (there is further discussion in Section 7.1.1). You should choose the forcefield carefully, based both on the system and on the question that you want to answer, as well as the simulation that you will be running. Here is a list of suggested questions you should consider when choosing a method or methods:

- What do you need to calculate?
- Do electrons need to be included?
- What levels of accuracy, simulation speed, and system size are required?
- Will single-point calculations give enough insight, or are dynamics required?
- Do you need to model the correct dynamics, or can you accelerate?
- How much sampling is needed for MD or MC to converge errors?
- What environment is needed? (e. g., solvent or not, boundary conditions)
- Do you need free energies?
- What parameters are available for fitted methods?

10.5.1 Molecular Mechanics and Forcefields

Molecular mechanics or forcefield methods are capable of modeling systems with millions or even billions of atoms, given the appropriate HPC facilities, and can be applied to timescales reaching nanoseconds routinely. They are widely used within biochemistry and materials science, and can be extremely effective when used properly. These methods are particularly useful for studying dynamical behavior over long distances or over long times.

As with any parameterized system, the key to reliability for forcefield-based calculations is the parameterization. Parameterizations have two key ingredients in their creation: first, the form of the parameterization and its flexibility; and second, the data to which the parameterization is fitted. A simple parameterization allows the essential interactions to be understood clearly, and avoids unexpected behavior from complex functional forms. However, it will not be able to model a wide range of behavior, particularly not beyond the original fitting data. A complex parameterization will allow a far wider range of behavior to be modeled, but runs the risk of over-complexity leading to unexpected behavior.

The data used to fit a parameterization will define its domain of applicability, and results beyond this domain should not be trusted. This is the perennial problem of extrapolation: going beyond the measured data is extremely risky. If a parameterization is not fitted to a set of interactions or to a particular behavior, then you should not apply it to this area. Thus, an important part of selecting a parameterization when judging its capability is to consider how it was developed, and what it was fitted to, and how well that overlaps with what you are modeling.

It is very hard for forcefields to model bond breaking and forming accurately, as these are inherently quantum mechanical processes. There have been developments in this area to allow forcefields which model the energy change during bond breaking (as discussed in Section 7.1), though these are never going to be as accurate as quantum mechanical methods. One approach to solving this problem is the use of quantum mechanics embedded into molecular mechanics (also known as QM/MM methods). These require considerable care and expertise to use, but have been applied to biochemical problems with some success. The question of boundaries and their effects on the simulation results is extremely important, and needs careful investigation in all circumstances.

Reliability

- Large systems: 10^5–10^9 atoms
- Long timescales: 10^{-9}–10^{-6} s
- Systems close to fitting
- Dynamical quantities
- Macroscopic quantities

Caution

- Bond making and breaking
- Systems far from fitting
- Processes involving electron rearrangement

10.5.2 Semiempirical Methods

The simplest way to include quantum mechanical effects into atomistic simulations is to use a semiempirical method. These methods are generally significantly faster than *ab initio* approaches, while allowing modeling of the electronic structure and the effects that this produces on interactions between atoms. These methods are sufficiently accurate to allow qualitative understanding in most cases, and quantitative understanding in some cases. The details of implementations are given in Section 8.6.

Semiempirical methods have at least two major weaknesses: the first is that they generally involve fitting to some experimental or *ab initio* data, and so have the same problem of a restricted domain of applicability as discussed for forcefield methods. This question of transferability is one which is encountered in many areas of atomistic computer simulation, and you must be aware of it, and test for its effects in your simulations as well as considering it when evaluating other people's work.

The second weakness in semiempirical methods is that they must make approximations in order to arrive at the simplified Hamiltonians that are used. These approximations will generally be controlled, but will have certain effects on the results generated, and there will be systems and simulations for which the approximations are good, and others for which they are poor.

Tight binding [1] is an approximation which can be derived from DFT, and has different levels of approximation. At the simplest level, the basis is assumed to be orthogonal, and the Hamiltonian is minimal. Even with this simple of an approximation, quantitative results can be achieved, and this method has been applied to simulations of many different systems, including semiconductors and metals. The basis for the Hamiltonian can be made larger and nonorthogonal, resulting in more complex and more accurate approaches. The end point for this process of adding complexity is the density-functional tight binding method, where the Hamiltonian matrix elements and their scaling are taken directly from DFT calculations of atoms overlapping, with some provision being made for self-consistency. Naturally, the accuracy that can be achieved with these methods is better than for simple tight binding models, though they require more computational effort.

Within quantum chemistry, expensive correlation calculations are often performed on top of semiempirical approaches such as the NDO family or extended Hückel theory. This is often the only way to estimate the relative importance of these contributions to the total energy and the properties of the system. However, as with tight binding above, care should be taken with the interpretation of results and their application.

Reliability

- Quantum mechanical effects (bond breaking/making)
- Qualitative and semiquantitative energies

- Medium systems: up to 10^3–10^6 atoms
- Moderate timescales: up to 10^{-12}–10^{-9} s

Caution

- Systems beyond fitting
- Quantitative energies: what was it fit to?

10.5.3 DFT

DFT is extremely successful in predictions of ground state properties, both of molecules and extended systems. With modern functionals, it is also reliable for chemical reaction barriers, and it is now used routinely throughout physics, chemistry, biochemistry, earth sciences and materials science. For calculations involving these problems, it is the method of choice.

However, it is not reliable in all systems. Any system with weak bonding (particularly van der Waals bonding) will not be well-described, and some care is needed with choice of functional when modeling breaking bonds. Strong correlations are very poorly described by DFT, and these simulations will normally produce poor results. Excited states are not well-described by DFT in general, which is a ground state theory. Thus band gaps and HOMO-LUMO gaps are often much too small.

There are recent developments which are relevant. The implementation of exchange within DFT (leading to hybrid functionals such as B3LYP and PBE0) improves band gaps and the description of partly filled orbitals, though with an increase in cost. Hybrid functionals for extended systems are often very expensive. These methods are also useful for correlated systems. The DFT+U method, where a local energy penalty is applied to bias occupancies of certain orbitals, is useful for transition metal systems, and is remarkably effective. Some care is needed, as the value of the parameter (known as U) can have a strong effect, and is effectively a fitting parameter. There have also been recent developments to add van der Waals energies and forces, both semiempirically and *ab initio*.

DFT simulations of a few hundred atoms are perfectly possible on a powerful workstation, and simulations of up to a few thousand atoms are possible on HPC facilities. Timescales of picoseconds are feasible, though timesteps of no more than 2–3 fs are required. Recent developments in linear scaling methods allow calculations on tens to hundreds of thousands of atoms, though these methods are still being characterized and optimized.

Reliability

- Ground state properties
- Quantitative energies and electronic structure
- Reaction barriers

- Moderate systems: up to 10^3 atoms
- Moderate timescales: up to 10^{-12} s

Caution

- Excited states
- Weak bonds
- Correlated systems
- Energy levels and gap sizes

10.5.4 Post-HF

Post-Hartree–Fock methods present a systematic route to calculating the correlation energy that is neglected by Hartree–Fock, and are often taken as the method of choice for calculations of total energies and energy barriers in small systems. However, they are computationally expensive, and the computational effort scales strongly with the number of atoms.

Perturbative approaches such as MP2 are very successful, but can only be used when the starting point is a good approximation to the problem. For problems such as bond breaking or where there are nearly degenerate energy levels, more complex methods will be needed. Methods in this area include configuration interaction and coupled cluster approaches. As the accuracy increases, the computational effort also increases, and the scaling of effort with system size often increases as well. The system size which can be addressed will be limited by these factors.

The methods use localized orbitals as basis sets, which are almost inevitably represented in terms of Gaussian functions. While there are basis sets which can be converged systematically, it is still very important to test the basis set carefully. Ensuring that the basis contains enough flexibility and the right kinds of functions to correctly describe the chemistry being modeled is vital.

There are significant efforts underway to improve the scaling of quantum chemistry calculations. Linear scaling MP2 calculations can be performed, allowing systems of up to fifty atoms to be addressed. Going beyond this will prove challenging, but these methods are being actively extended. The alternative, as mentioned above, is to use a semiempirical method for the Hamiltonian, and to build post-HF calculations on top of that. These methods are not used for molecular dynamics, and geometries are normally found by using less computationally expensive methods.

Reliability

- Energy levels
- Reaction barriers

- Small systems: from 5–50 atoms
- Correlation energy

Caution

- Basis set size
- Sufficient accuracy
- Model size

10.5.5 Post-DFT

The term post-DFT encompasses a number of different areas: many-body perturbation theory approaches such as GW and BSE; time-dependent DFT; and quantum Monte Carlo. These are complementary approaches, which are normally used for different problems.

The perturbation theory approaches have the same limitation we discussed for MP2: they rely on the starting electronic structure being a good approximation to the final result. The input for GW is a set of single particle bands, which are taken to be an starting approximation for the quasiparticle bands. While there are active research programs around self-consistent GW approaches, it is important to check that these input bands capture the essential physics of the system. In particular, if using DFT, the gap should not be closed. The BSE has similar requirements, and is computationally more intensive than GW.

TDDFT is increasingly widely used, particularly for molecules. The energy levels generated are significantly more reliable than DFT levels, though as with DFT there are questions over how to construct accurate functionals. The TDDFT functionals are less well-explored than DFT, and some care is needed. One well-known problem area for TDDFT is in long-range charge transfer excitations, which are very poorly described. Most TDDFT calculations are performed in the linear response regime, and this adds an extra approximation for the exchange-correlation kernel. Use of TDDFT with periodic boundary conditions and bulk systems requires considerable care to account for any polarization effects which might arise.

QMC is becoming more widely used, but is still within the realm of expert practitioners. Its main use is calculating energies, and it is often applied where energy differences are small or in problems which involve effects that DFT does not model well. As with all statistical methods, QMC relies on convergence of the computed properties with sampling. It also requires input wavefunctions, which are often taken from DFT or Hartree–Fock calculations. Reliable forces are not yet available from QMC methods, though there is active research in this area.

Reliability

- Band structure
- Optical transitions

- Excited states
- Energy differences
- Subtle interactions, for example, dispersion

Caution

- Input bands (will DFT provide a good input?)
- Charge transfer
- Forces (especially with QMC)

10.6 Writing About the Simulation

There are two main aims when you are writing up your simulation results: the first aim, which is the most obvious and most important, is to communicate the scientific importance of the results; the second is to enable others to understand *how* you have achieved these results, and to reproduce them if necessary. It is the second of these points that we will discuss here.

What is the minimum amount of information that would be needed to reproduce your simulation? The following list is what should be available to other researchers if requested, or deposited as supplementary information:

- Input coordinates
- Input parameters (or input file)
- The name and version of the simulation code used (and possibly any source code you have written for the simulation)
- The platform used to run the simulation, and the compiler(s) used
- Any further inputs, for example, forcefield variants, pseudopotentials (or how they were generated)
- Details of the simulation protocol (e. g., equilibration, annealing etc.)

It is important to understand that results will vary to some extent or other between the same code run on different platforms (or computers) and even between the same code from different compilers on the same platform. Most of these differences are in the noise, but there can be significant errors, and publishing at least the code version, platform and compiler will enable some form of reproducibility.

There is a new approach emerging within computational simulation, not specifically atomistic simulations, known as *reproducible research* [2–4], which aims to make simulations reproducible. The essential argument behind this approach is that it should be possible to reproduce any given piece of computational research. For this to be possible, both the input data and the computational code must be available. Many scientists are unwilling to publish their code, either because it may be commercially valuable or because it is inelegant or poorly written. But code does not necessarily need to be in a good state to be published [5]: it is enough that it is available. The question of commercial exploitation is a harder one, and also applies to the use of commercial codes: this is why it is vital to give details of the version of any commercial code used (functionality, default settings and even core algorithms can change between versions).

You should document the details of each run you perform, and make sure that you keep backed-up copies of both input and important output files. This seems like obvious advice, but will avoid at least two traps. First, the danger of running all simulations in one directory and rerunning and erasing a previous result. It is far too easy to run a calculation, find some answer, and then overwrite it by changing an input flag and rerunning. If you want to run a simulation building on the output of a previous simulation, make sure that you have a copy, and possibly establish a new directory before running. Second, the danger of running calculations in some form of scratch directory, standard in many HPC environments, which is not backed-up is that you will lose results. We have known people lose several months work through poor backing up and documentation of simulations.

You should also analyze, visualize, document and write up your results as you progress through simulations. There is a temptation to run calculations, look at the important points, and move to the next problem, with the intention to write up when a complete project has been finished. However, it will be much harder to document the simulation protocol you have used several months after using it. Moreover, it may be hard to remember exactly why you chose a particular series of calculations, or what a particular directory was used for. Systematic documentation and naming will avoid these problems. Several of the reproducible research approaches allow documentation and analysis within one package, and can aid the process.

10.7 Checklists

Framing the questions

- What is known experimentally?
- What has already been calculated?
- What can be calculated and what is needed?

Which method will answer which question?

- Are electrons needed?
- What dynamics are necessary?
- How many methods will be used?
- Does the input of one method depend on the results from another?
- Do I understand the method(s)?

Resources

- Time
- CPU power
- Total CPU time
- Methods and codes
- Computational support
- Money (to pay for codes and CPU time)

Model

- Is my model large enough (boundary conditions)?
- Is my model too large (computational efficiency)?
- Where can I obtain the coordinates?

Calculation sequence

- Have I characterized the system?
- What are the important energy, length and timescales?
- What convergence criteria do I need at each stage?
- What simulation cell size do I need at each stage?
- Is my system equilibrated correctly?

Further Reading

Gawande, A. (2011) *The Checklist Manifesto*, Profile Books.

A description of how checklists are used in many technical areas, concentrating on surgery. An excellent introduction to how and why checklists are used.

References

1 Goringe, C.M., Bowler, D.R., and Hernández, E. (1997) Tight binding modelling of materials. *Rep. Prog. Phys.*, **60**, 1447–1512.

2 Fomel, S. and Claerbout, J.F. (2009) Reproducible research. *Comput. Sci.Eng.*, **11**, 5–7.

3 Mesirov, J.P. (2010) Accessible reproducible research. *Science*, **327** (5964), 415–416.

4 Peng, R.D. (2011) Reproducible research in computational science. *Science*, **334** (6060), 1226–1227.

5 Barnes, N. (2010) Publish your computer code: It is good enough. *Nature*, **467**, 753.