Review

Author for correspondence:
Peter V. Coveney
e-mail: p.v.coveney@ucl.ac.uk

Uncertainty quantification in classical molecular dynamics

Shunzhou Wan[1], Robert C. Sinclair[1] and
Peter V. Coveney[1,2]

[1]Centre for Computational Science, University College London, Gordon Street, London WC1H 0AJ, UK
[2]Institute for Informatics, Science Park 904, University of Amsterdam, 1098 XH Amsterdam, The Netherlands

SW, 0000-0001-7192-1999; PVC, 0000-0002-8787-7256

Molecular dynamics simulation is now a widespread approach for understanding complex systems on the atomistic scale. It finds applications from physics and chemistry to engineering, life and medical science. In the last decade, the approach has begun to advance from being a computer-based means of rationalizing experimental observations to producing apparently credible predictions for a number of real-world applications within industrial sectors such as advanced materials and drug discovery. However, key aspects concerning the reproducibility of the method have not kept pace with the speed of its uptake in the scientific community. Here, we present a discussion of uncertainty quantification for molecular dynamics simulation designed to endow the method with better error estimates that will enable it to be used to report actionable results. The approach adopted is a standard one in the field of uncertainty quantification, namely using ensemble methods, in which a sufficiently large number of replicas are run concurrently, from which reliable statistics can be extracted. Indeed, because molecular dynamics is intrinsically chaotic, the need to use ensemble methods is fundamental and holds regardless of the duration of the simulations performed. We discuss the approach and illustrate it in a range of applications from materials science to ligand–protein binding free energy estimation.

This article is part of the theme issue 'Reliability and reproducibility in computational science: implementing verification, validation and uncertainty quantification *in silico*'.

1. Introduction

Computational methods offer a route to understand and predict the structure, dynamics and thermodynamics of molecular systems. A large fraction of these are primarily based on molecular dynamics (MD) simulations, first developed in the late 1950s [1]. These methods, however, find their applications only in a limited segment of industry. Only in the past decade have they started to be used in anything approaching an actionable context, for example in materials manufacturing [2] and in the pharmaceutical industry [3]. Their main use is as a form of post hoc rationalisation of experimental findings. In comparison to macroscopic modelling methods such as the finite-element method (FEM) or computational fluid dynamics (CFD), which enjoy widespread use in engineering applications, MD methods remain primarily confined to academic research, largely due to their lack of reproducibility and limited accuracy, as well as the frequent long duration and computational expense required for their use. Speed, accuracy, precision and reproducibility are essential in any method which is to be relied upon for taking actionable decisions and thus to become valuable in diverse applications, including *inter alia* industrial and clinical contexts. For that, we need uncertainty quantification (UQ), verification and validation (V&V) or VVUQ. But while careful control of uncertainty is the mainstay of weather forecasting, along with many branches of engineering and applied mathematics, it is rather rarely performed in disciplines such as physics and chemistry where much time is spent investigating matter at shorter length and time scales than the macroscopic ones of direct concern in many real-world situations.

The purpose of the present paper is to assess the use of uncertainty quantification in the field of molecular dynamics simulation and its relationship to the reproducibility of the method. We do not provide a comprehensive review of molecular dynamics applications. We shall spend much less time looking at verification and validation, which are respectively concerned with whether molecular dynamics computer programs are solving the correct equations and how well their output agrees with experimental observations [4].

Computational predictions are increasingly being used to predict outcomes and provide recommendations in a variety of industrial and policy-making contexts. One of the central aims of UQ is to facilitate decision making [5]; this is enabled by providing probabilistic statements about quantities of interest, in a timely fashion, and helping decision makers to decide what actions to take that maximize the likelihood of the desired outcome. Ensemble methods for probabilistic weather prediction have become a routine part of the practice of weather forecasting [6]. Simulation results have been used to help governments and non-governmental organizations to make decisions as to how to help plan ahead for mass movements of refugees [7]. Accurate and rapid binding free energy predictions based on classical molecular dynamics [8] are starting to impact decision making in the pharmaceutical industry and clinical settings [9,10]. A very recent and topical example is predictions emanating from simulations of an epidemiological model by Ferguson *et al*. [11–13] which were used to guide UK government policy in addressing the COVID-19 pandemic.

The steady increase of computational power permits investigations of a large diversity of models over increasing length and time scales. This makes it even more essential to systematically assess the reliability and reproducibility of the methods used and the results generated, as such large-scale applications involve a vast amount of computational time and human effort. Uncertainty quantification, along with validation and verification, ensures that the results are reliable and reproducible while conferring greater confidence on predicted outcomes. Without UQ, the usefulness and value of simulations are diminished, and the confidence in computational results degraded. Science and engineering manifestly advance faster when there is less time wasted on pursuing false leads.

Given the increasing application of scientific computation in critical decision making, uncertainty quantification has been the subject of broader attention in a range of domains [14]. In computational physics and computational chemistry, for example, there are well-documented cases in the literature assessing the reliability with which various quantities of interest can be computed. In ensembles of exchange-correlation functionals for density-functional theory calculations, large error bars have been observed on energy differences at different states [15], in reaction energies of specific chemical systems [16], within reaction kinetics of surface transition states [17], and estimated free energies of activation [18]. Uncertainties in a model's parameters contribute to the uncertainties in its predictions. Careful and systematic uncertainty quantification applied to the parameters enables one to estimate prediction uncertainty based on parameter uncertainty inflation or deflation [13,19]. In some instances, prediction errors due to model inadequacy can be handled by statistical correction of predictions, which may provide a reliable uncertainty measure [20]. Various surrogate methods have been developed for the estimation of prediction uncertainty, such as bootstrap-based methods, Gaussian process regression, neural networks and deep learning ensembles [21–23]. Gaussian process regression has been employed to identify particular calculations within a given dataset for which the uncertainties exceed a given threshold [24,25]. The data points flagged up can be investigated with refined models and more accurate methods that may be added back to the dataset if their uncertainties then fall below the threshold concerned. Such validation and updating can improve the models and enhance the quality of subsequent predictions [24].

One major application of MD simulation that we consider is the prediction of the binding affinity of a lead compound or drug candidate with a protein target, of major relevance in drug discovery and personalized medicine. That target may be respectively either a generic protein or a sequence-specific variant, reflecting the fact that individuals respond differently to a given drug based on their genetic make-up. The binding affinity, also known as the free energy of binding, is the single most important initial indicator of drug potency, and the most challenging to predict [26,27]. Another case we look at here is how one seeks to make actionable predictions in support of advanced materials discovery. The guiding principle in all instances is to seek to make reproducible predictions.

Reproducibility is an intrinsic feature of the scientific method, whether experimental or computational. Scientific methods should yield the same results in a statistical sense regardless of who performs them. Indeed, the lack of reproducible results in the published literature is of current concern in the wider scientific community [28–30]. It should be noted that chaotic dynamical systems exhibit extreme sensitivity to initial conditions, making accurate predictions impossible and one-off observations largely unreproducible even though their underlying dynamics is deterministic. Molecular dynamics is a case in point for which these issues need to be addressed. To restore the predictive power of the scientific method to such systems scientists contend that, while the accuracy with which we can simulate an individual chaotic process is severely limited, accuracy in an averaged statistical and reproducible sense may still be possible [31]. The purpose of the present paper is to assess the reproducibility and intrinsic uncertainty of molecular dynamics simulation. We illustrate the issues by way of some examples drawn from materials and life sciences. The discussion is, of course, applicable to all areas of classical molecular dynamics.

2. Sources of error in classical molecular dynamics

There are two sources of error accruing in MD simulations, due to systematic and random sources. In order to get a full grip on uncertainty in MD simulations, one needs to be able to identify both. Systematic errors originate in things like the imperfect design, parameterization, conduct and/or analysis of a study, which result in an estimate of a property deviating consistently from its true value. Random variation—also called system noise, aleatoric or stochastic error—on the other hand, is caused by the intrinsically chaotic nature of classical molecular dynamics and produces apparently random deviations from the notionally 'true' value of an observable. It should be

noted that in some cases no consensus definitions for the 'true value' may exist; more precisely, differences in definitions from modelling and experimental studies can then contribute to observed errors in validation studies. The glass transition temperature and the setting of cements, for example, have a number of different operational definitions although only some of them are endorsed as accepted scientific or engineering standards. The simulations should proceed from a statistical-mechanical ensemble corresponding to the experimental conditions, and properties calculated from expectation values may then be compared with their corresponding experimental counterparts. Quantifying systematic errors requires first bringing the random components contributing to the errors under control.

(a) Systematic errors

Systematic errors are introduced by inaccuracies inherent to the system investigated and within the measurement method performed. They come from the assumptions and approximations made when a theory is applied, a model is constructed, or a process is mimicked by the simulation of a real-life problem. In constructing a model, there are many choices to be made, including: which degrees of freedom are to be modelled explicitly, what components are to be excluded, the kind of interactions between the components, what boundary conditions are to be used, and so on. As Michael Levitt has stated: 'the art is to find an approximation simple enough to be computable, but not so simple that you lose the useful detail' [32]. In principle, a higher level of resolution should produce more accurate predictions than a lower level one, although in practice it is not always the case because of the quality of the theory employed, a fortuitous cancellation of errors, or the way that the methods are implemented which may not be fully verified. It is not uncommon that simple methods outperform more complicated methods in the simulation field, as has been seen in the blind SAMPL free energy prediction competitions [33–35]. In drug discovery approaches, a ligand–protein model with explicit water molecules is usually better than one with implicit water. These choices all affect the outcome of a simulation, usually in a deterministic way. Biases in the interaction parameters chosen to represent the system can significantly influence the results; for example, different protein force fields favour different secondary structure types [36,37], populating either helical or sheet-like structures within independent simulations. When the cause of such systematic errors can be identified, it can be reduced or even eliminated, as shown for example in recent simulations with state-of-the-art force fields [38].

The implementation of the model in an appropriate MD engine and the calibration of the engine can also influence the results. The thermodynamic conditions, such as constant volume or pressure in a closed system, must be specified. Multiple factors need to be carefully considered in the preparation of the molecular models, such as choice of force field, protonation and tautomeric states, buffer conditions, use of restraints and constraints, thermostat and barostat, free energy estimator and finite machine representation of floating point numbers [39,40]. A few operational parameters need to be fine-tuned, including those for temperature and pressure couplings, for the calculation of long-range interactions, for the time step(s) used within the integration algorithms, and so on. Other factors, such as the introduction of numerical integrators and the accumulation of rounding error [31], may also lead to systematic patterns of error. Moreover, it is entirely possible that molecular dynamics may manifest a pathology which we recently discovered in the simulation of simple chaotic systems described by the generalised Bernoulli map on digital computers [41]. It is caused by the limitations of the IEEE floating point numbers in describing the statistical behaviour of systems with such exquisite sensitivity: for certain parameter values, ensemble and long-time averages, designed to address random errors, also contribute substantial systematic errors to predicted properties. For those particular parameter values, the pathology cannot be mitigated by increasing the precision of these numbers. While single precision floating point numbers are inadequate for such extremely simple chaotic systems, double precision floats can produce much better behaviour for most other parameter values (Milan Kloewer, private communication with PVC). However, this leaves open the question as to whether double precision floating point numbers are themselves sufficient to handle the complexity of real world

molecular dynamics and fluid turbulence, which originate in dynamical systems that are many orders of magnitude more complicated.

(b) Random errors

Given the extreme sensitivity of Newtonian dynamics to initial conditions, two independent MD simulations will sample the microscopic states with different probabilities no matter how close the initial conditions used [42]. The difference thus produced in two simulations introduces a variation in results that can often be larger than the quantity of interest, making the results practically useless. Large hysteresis can be observed in cases where adequate sampling of all relevant conformational substrates is not achieved [43,44]. It should be noted that a seemingly low standard deviation does not guarantee convergence; it can appear when simulations remain trapped in a single energy well [45].

The impact of the chaotic nature of molecular dynamics has not been widely recognized in the molecular dynamics field. Most accounts give surprisingly short shrift to it, a notable exception being the recent book by Leimkuhler and Matthews [31], albeit it does not address the issue of uncertainty quantification and the estimation of thermodynamic quantities. Extensive studies we have performed in recent years show that molecular dynamics systems indeed exhibit extreme sensitivity to initial conditions [46–49]. From our investigations, we observe that the properties one computes from molecular dynamics trajectories appear *superficially* to be described by a Gaussian random process (GRP) with a normal distribution denoted by $N(\mu, \sigma^2)$, characterized by average μ and standard deviation σ (the square root of the variance in the data). Note, however, that a normal distribution cannot be assumed and in fact there are frequently significant deviations from such statistics in nonlinear dynamical systems of which molecular dynamics is an excellent example [50].

In a recent study, for example, the effect of box size on simulations of protein dynamics in water was reported [51]; the authors reported that calculated values of various properties, such as the stabilities of the unliganded and liganded states of human haemoglobin and the density or number of hydrogen bonds per water molecule, changed systematically with an increase in box size; the authors maintained that a surprisingly large box of 150 Å was required to obtain meaningful results. Although at first sight, this dependency on the box size appears to be an example of a systematic error in the simulation, it is in fact caused by a lack of reproducibility in the study which becomes manifest when random errors are taken into account [52]. Indeed, the ensuing debate [51–54] highlights the importance of setting up systems correctly for simulation and, more importantly, applying ensemble approaches to get statistically significant results. The issue at stake is also important for the evaluation of the reliability of scientific research, as is clear from the follow up studies by the debaters [45,55]. As we noted above, without first handling the stochastic errors, it is not possible to assess correctly the nature and magnitude of the systematic errors, and to interpret findings convincingly.

3. Uncertainty quantification in molecular dynamics simulations

Although it was recognized more than two decades ago that one-off classical molecular dynamics simulations do not generate consistent protein conformations [56,57], systematic investigation as to how to make these calculations reproducible had not been performed until recently. Considerable effort has been invested in the development of so-called 'enhanced sampling protocols' in order to accelerate phase space sampling, their purpose being to make computed properties more reliable by demonstrating more rapid 'convergence' of computed properties. These enhanced sampling protocols accelerate molecular dynamics to overcome high energy barriers using methods such as bias potential approaches [58,59], Markov models [60], orthogonal space random walk [61], self-guided Langevin dynamics [62] and Hamiltonian replica exchange [63]. However, in all these cases, it is quite impossible to calculate the (equilibrium or other) probability distribution function from one-off simulations, against which expectation values

would be calculated; instead, expectation values of various observables are reported. Ensembles of such enhanced sampling simulations show that there is a significant variance between the expectation values computed from individual replicas [64].

Indeed, ensemble-averaging is not just a practical consideration invoked in the repertoire of uncertainty quantification methods. When molecular dynamics is used, as it frequently is, to estimate thermodynamic properties, such as the free energy of a system, it should be recalled that the connection between microstates (generated by individual MD simulation trajectories) and thermodynamic properties is achieved using ensemble averages. This is true whether the system is in or out of equilibrium. The very common resort to perform so-called 'long-time averages' of a single microsystem appeals to the ergodic theorem, which is in fact only valid for long times at which this time average should converge to the ensemble average. In reality, that time interval would need to be of the order of a Poincaré recurrence time—a truly astronomical epoch—for the equality to hold. In practice, it is taken to be as long as authors deem to be reasonable, and, compounding this, we must face the fact, mentioned below, that the accuracy of these long-duration trajectories is severely limited.

In the ergodic hierarchy of dynamical systems, those which approach and reach equilibrium must be at least mixing [42]. Mixing implies ergodicity, but the converse is not true. Mixing systems exhibit the tell-tale property of dynamical chaos: neighbouring trajectories, no matter how close, diverge exponentially, at a rate given by a Lyapounov exponent. The point is that we are dealing with two levels of description and those wedded to trajectories and Newtonian mechanics think only in terms of one, the trajectories which follow Newton's equations of motion. To understand the concept of equilibrium, one must work with probability distributions (computed from ensembles) which obey Liouville's equation. Ergodic theory is about the large-scale probabilistic properties of dynamical systems, including in particular their long-time behaviour. The condition for a solution of the Liouville equation to reach an equilibrium state is that the dynamics must be mixing [42]; at the level of trajectories, it means that neighbouring trajectories diverge exponentially. An analogy may help. Imagine some perfume in a bottle in one corner of a room. The cap is removed and the vapours suffuse the room. Equilibrium is reached when the density of the perfume is uniform throughout the room—that is the equivalent to phase space equilibrium, when the probability distribution function no longer changes with time.

Since we can never know the true initial conditions for a real system (which arise as a consequence of whatever it was doing before we started to observe it), we are obliged to formulate the approach to equilibrium in probabilistic terms. Indeed, even a single trajectory associated with a given initial condition becomes increasingly inaccurate as time passes, since the exquisite sensitivity of the dynamics means that round-off errors accruing during the time integration of the equations of motion inevitably put the system on orbits other than the one it began on.

The lack of reproducibility thus stems primarily from the intrinsically chaotic nature of classical molecular dynamics. Other sources of uncertainty may play a role, many being potentially tractable; these include: the theory or the model used, the extent of convergence of the numerical method, the reliability of the software (which may not have been verified), the way the software is used and so on [8,39]. We therefore focus on ensemble-averaging, which is mandatory in statistical mechanics, for the convergence, reproducibility, reliability and uncertainty quantification of properties obtained from MD simulations. If we adjudge the system to be in a state of equilibrium, we can in addition perform time averaging, a procedure generally bereft of meaning out of equilibrium.

(a) Ensemble method

Extensive studies we and others have performed in recent years [8,44,46–49,64–74] confirm that the most effective and reliable computational route to reproducible binding free energies of ligands to proteins using MD simulation can be achieved using ensemble methods. The same conclusion has been drawn from MD simulations in other areas, including studies on materials applications such as graphene-based systems [75], on DNA nanopores using coarse-grained MD

simulations (Ahmad K, Coveney PV. 2019 unpublished work), on rate parameter estimation for binding kinetics [76] and so on [77,78].

An ensemble approach employs a set of independent MD simulations, referred to as 'replicas' both in statistical mechanics and within the uncertainty quantification domain, to obtain the required averages and associated. The key feature of such simulations is the use of ensembles and—for systems at equilibrium—time averaging. It is useful to recognize the stochastic nature of these simulations; it can be convenient to approximate the statistical properties of such ensembles as Gaussian random processes [42]. The requirement for the number of replicas and the temporal duration of the simulations depends mainly on the property of interest and the conformational space being sampled. Each replica needs to be simulated for long enough to sample the most relevant conformational space in order to evaluate one or more properties of interest. For free energy calculations using an endpoint approach, for example, the stable binding states are the most relevant. There is no theoretical means to establish the number of replicas required to produce low errors from ensemble simulation: the criterion for ensuring convergence of the ensemble average is to establish the number N of replicas required such that using $N+1$ of them makes no significant difference to the expectation values calculated. Of course, this can also be looked at another way, as amounting to a trade-off between the amount of computation one performs (which increases linearly with N) and the size of the error one is willing to tolerate (which reduces roughly as $N^{-1/2}$). The computational costs can be optimized when a cost-to-accuracy ratio is carefully considered in a situation-specific and task-specific manner. Our studies show that starting from good initial structures, accurate and reproducible results can often be achieved from an ensemble simulation consisting of 25 replicas with 4 nanoseconds production runs each using endpoint approaches [8]. For a drug screening project, where a large number of compounds need to be evaluated, a coarse-grained workflow with a smaller number of replicas can indeed be useful to distinguish binders from non-binders [79]. In such cases, replicas differ only in terms of the random number seeds selected to assign initial atom velocities. In other situations, it may additionally be necessary to randomly vary the initial atomic configuration, that is the spatial coordinates, as well [80]. These matters are discussed further below.

In general terms, the principles discussed above are applicable to all-atom MD and coarse-grained MD. The reduction in the number of degrees of freedom achieved by grouping several atoms into single particles or pseudo-atoms, as is done in coarse-grained MD, reduces the level of fluctuations in such systems. While this form of coarsening of the model's representation can typically lead to a decrease in accuracy, the benefits which accrue are an ability to study larger systems for longer time periods; the reduced degrees of freedom also reduce the phase space that needs to be sampled. We find that smaller ensembles are required for coarse-grained MD than all-atom MD. This typically leads to an ensemble with around two-thirds the number of replicas as compared to all-atom MD ensembles, as we have shown in recent work with graphene oxide dispersions [81] and DNA nanopores in lipid bilayers [82].

(b) Performing ensemble simulation

The requirement to simply run a large ensemble of replicas may sound trivial but it comes with significant overheads in terms of managing the execution of simulations and collation of output data. It will be clear that using ensemble methods greatly adds to the computational cost of a study and to the wall clock time unless one has access to modern high-performance computers which are equipped with large numbers of nodes, cores and often accelerators. In such cases, in the time it takes to run one simulation, one can produce the output for all of them. There is then a lot more data to handle and process. The key step from the overall technical perspective is to bring all these output trajectory data together and then perform the analysis on the aggregate of all of it. This can be done in a number of ways, but the one we generally use is called a bootstrap error [83]. Given N results from an ensemble of simulations, the bootstrap method involves calculating the distribution of means from resamples of size N from the original data. Many resamples are taken, typically greater than 10,000, with replacement. If the original sample is representative of the

true distribution, this method can provide error bounds or confidence intervals on any calculated value. The bootstrap error behaves similarly to a standard error; indeed, it is meaningful for quantities that have non-Gaussian distributions. This is of practical value in cases where we do not know the distribution of the quantity of interest.

Evidently, effective automation is necessary to manage the extra effort involved, and efficient sampling techniques are required to make these kinds of workflow possible. We have previously developed software to assist us in this task for the computation of binding free energies, the so-called 'binding affinity calculator' [84]. This in turn led us to develop software for more general forms of uncertainty quantification, and to extend this to address verification and validation too. In particular, we have developed the VECMA Toolkit [30,85], as an open-source, open development project which enables us to apply these methods much more widely, to address uncertainty quantification in a set of diverse domains. For example, the toolkit includes a python library called EasyVVUQ, which permits users to instrument their own codes with capabilities to perform UQ using a wide range of methods, including quasi-Monte Carlo (the method described here), stochastic collocation and polynomial chaos [29,86].

4. Statistical distributions revealed by ensemble simulation

The use of ensembles in molecular dynamics simulation only started to be systematically and routinely viable since the advent of the petascale era (that is, in a little over the past ten years), as a result of the vast increase in the number of nodes, cores and accelerators available on supercomputers. An instructive thing to do is to plot the frequency distribution of observables as it emerges from all the members of an ensemble, as this gives us an indication of the nature of the distributions we can expect. Figure 1 shows a set of examples of the data which typically emerges from these studies, and is drawn from work we have performed over the past few months [79].

It is clear from these plots that, while they are approximately Gaussian, they exhibit deviations from the standard bell curve expected on the basis that the variables are independent of one another, as one assumes in conventional statistics. Instead, we find that the distributions have a skewness associated with them, the asymmetry favouring the occurrence of values of the observable higher than the mean. The majority of the distributions have positive excess kurtosis, meaning they are heavy-tailed relative to a normal distribution. Our predictions of non-normal distributions prompted an investigation of the distributions of experimental binding free energies from a large number of independent measurements over time: these also exhibit non-normal properties (Ian Wall and Alan Graves, private communication, 2020). Caution is therefore required when statistical comparisons are made between predicted and measured thermodynamic properties, as it is widely assumed that these data are normally distributed. Such behaviour is at first sight unexpected, until it is recognized that these systems all display chaotic behaviour as well as long-range interactions [31]. A Gaussian distribution of free energy results can only be assumed for harmonic systems or transformations that can be approximated by linear response theory (see, e.g. Hummer *et al.* [87], Shirts and Pande [88], or König *et al.* [89]). Most free energy calculations are strongly influenced by anharmonic terms (e.g. van der Waals interactions), are not performed in a homogeneous environment (e.g. in a protein), or exhibit more than one dominant conformational substrate [44]. The underlying nonlinearities in the dynamics are what account for both the presence of chaos and non-Gaussian statistics. The phenomenon is well known in turbulence: there it is caused by very long-range hydrodynamic interactions mediated by energy dissipation. It is not anticipated from our experience of studying linear systems with short-range interactions at equilibrium.

That the equilibrium distributions arising in molecular dynamics should be non-Gaussian may appear surprising, given the accounts in most textbooks and lectures, which transfer themselves into research articles very readily. The reason for the presence of non-normal statistics in such systems at equilibrium comes from the fact that here too we are dealing with the infinite range interactions mediated by Coulomb forces. In a computer simulation of a closed system, such as the canonical or (N,V,T) ensemble, the molecular dynamics is driven by the existence of thermostats
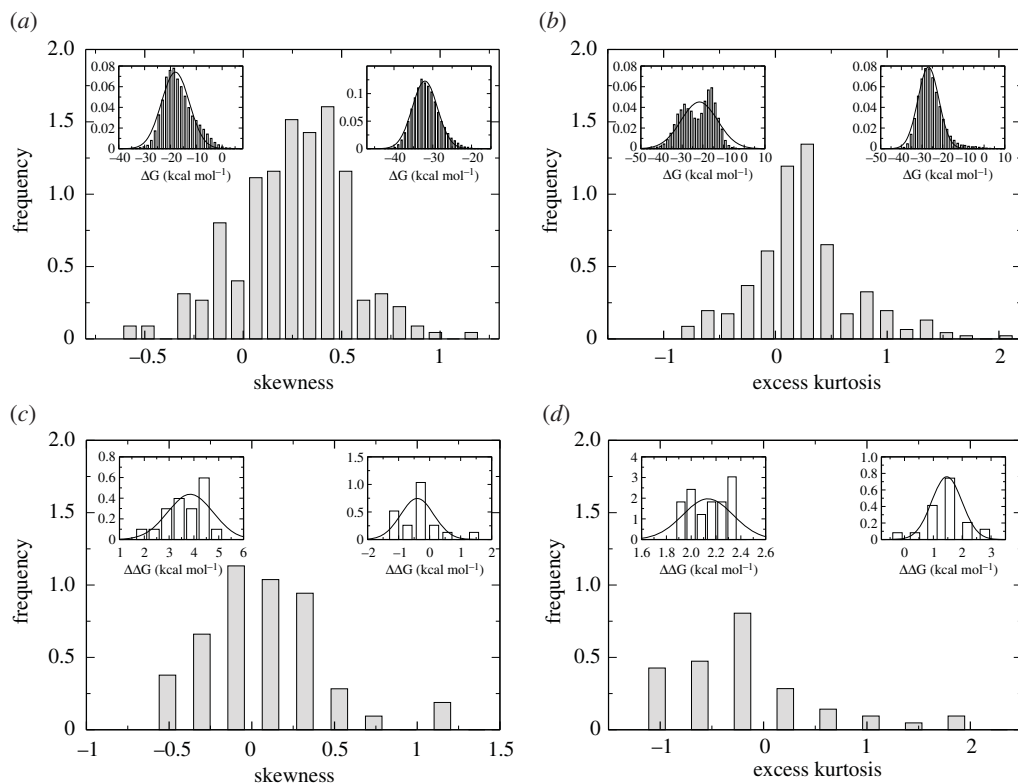
Figure 1. Molecular dynamics equilibrium distributions with long-range interactions are non-Gaussian. The main figures display the Fisher–Pearson coefficient of skewness (*a,c*) and the excess kurtosis (*b,d*) for distributions of predicted binding free energies (ΔG), or binding free energy differences ($\Delta\Delta$G), using the ensemble-based molecular mechanics Poisson–Boltzmann surface area approach (ESMACS) (*a,b*) and thermodynamics integration (TIES) (*c,d*) approaches, respectively. The ESMACS results are obtained from 250 ligand–protein complexes, each with 25,000 frames accumulated from an ensemble simulation with 25 independent replicas. The TIES results include alchemical transformations of 50 pairs of ligands, from ensemble simulations comprising 20 or 40 replicas each. The inset shows distributions of binding free energies for two ligands, or ligand pairs, with the most negative and most positive skewnesses or kurtoses respectively. The best-fit Gaussian distributions are shown by black solid lines. See also Wan *et al.* [8].

(and barostats in e.g. the (N,p,T) ensemble). The dissipation of energy within the system causes long-range correlations to be set up, which manifest themselves in the non-Gaussian nature of the statistics.

The practical implications of this discovery are important to apprehend. Non-normal distributions imply the occurrence of more 'outliers' and an increase in the observation of so-called 'rare events', making it harder to infer poor agreement between theory and experiment without far more data to pin down average behaviour. Naïve interpretation of correlation plots as implying poor agreement between experiment and theory when data do not magically cluster close to the 45-degree line need closer assessment; this is discussed further by Wan *et al.* [8].

5. Biomedical simulation

There are various approaches to estimate the magnitude of the binding free energy (a measure of how strong the interaction is between a ligand and its target protein), based on different theories and approximations [90]. The 'informatics' based approaches are, in the current era, usually the output of docking studies in combination with so-called 'machine learning' [91]. The linear interaction energy (LIE) method [92] is an approximation of linear response theory; molecular

mechanics Poisson–Boltzmann surface area (MMPBSA) and molecular mechanics generalized Born surface area (MMGBSA) methods [93] are based on invoking a continuum approximation for the aqueous solvent to approximate electrostatic interactions following all-atom molecular dynamics simulations, and, finally, so-called 'alchemical' methods, including thermodynamic integration (TI) and free energy perturbation (FEP), have a theoretically firm foundation, the former being exact and the latter an approximation. In practice, relative alchemical free energy methods have a restricted domain of applicability due to sampling requirements for convergence and, usually, the need to keep electrostatic charges constant during alchemical transmutations. Absolute binding free energies derived from alchemical methods are at least an order of magnitude more expensive. Binding free energy can also be calculated using approaches based on potential of mean force; such calculations usually employ enhanced sampling approaches such as metadynamics, adaptive biasing forces or umbrella sampling [94]. The choice of which computational method to use is influenced by the desired accuracy, precision, time to solution, computational resources available and so on.

(a) Ensemble method for endpoint approach

Endpoint free energy methods allow one to explore configurational space in the protein–ligand bound and unbound states only, providing an efficient and accurate approach for the calculation of state functions such as free energies. MMPBSA and MMGBSA approaches are two commonly used endpoint free energy methods which require direct simulation of the two physical states. To generate the structures of these states, one can use a 3-trajectory (3-traj) approach in which separate MD simulations are performed for the ligand, apo (free) protein and ligand–protein complex. Alternatively, one can use a 1-trajectory (1-traj) approach in which a single simulation is performed for the complex; a 2-trajectory (2-traj) variant allows for flexibility in the complex and one of the other two. The conformations for the ligand, protein and complex all being extracted from the complex simulation, the 1-traj approach makes use of an assumption that the conformations of the separated ligand and protein are similar to those of the complex. The assumption in the 1-traj approach is based on a lock-and-key hypothesis in which a substrate fits perfectly into the active site of an enzyme just like a key fitting into its lock. In the 1-traj approach, noise is significantly reduced as the energy terms are largely cancelled out. That assumption is questionable in many cases, however, as binding typically leads to conformational changes in both protein and ligands. The 3-traj approach does not use this assumption but the amount of noise is substantially increased when taking energy differences from three individual simulations. It is important to recognize that the 2- and 3-trajectory variants only become possible when ensemble-based methods are used, as extensive averaging is required to reduce the fluctuations present in individual trajectories, as is discussed further below.

(b) Errors in endpoint approaches

The variation in the 1-traj free energy calculations based on ensemble simulations was investigated systematically by Sadiq *et al*. [46] and by Genhenden and Ryde [93] using MMPBSA and MMGBSA methods, respectively in 2010. The estimated free energies from two independent MMPBSA calculations of the same molecular system can vary by more than 10 kcal/mol in smaller molecule–protein complexes [47,49,93], and by up to 43 kcal/mol in larger and/or more flexible ligands bound to a protein such as peptide–MHC (major histocompatibility complex) systems [48]. With the ensemble method, however, a meaningful ranking of binding free energies is generated. The 3-traj approach is able to address the role of the adaptation energy—the free energy associated with the conformational changes upon binding. The large adaptation energy, up to 39 kcal/mol, indicates that it is necessary to invoke the 3-traj approach for flexible small-molecule–protein binding. This has been confirmed by several subsequent studies [64,65,68] which show that incorporating the flexibility of the receptor and ligand improves the prediction of

binding free energy ranking. There are also cases where incorporating flexibility does not improve the ranking [67], indicating that binding is mediated by a lock-and-key mechanism.

(c) Ensemble endpoint simulations

To generate reliable, precise and reproducible binding free energies from MMPBSA and MMGBSA approaches, we have proposed an ensemble-based MD approach, named 'enhanced sampling of molecular dynamics with the approximation of continuum solvent (ESMACS)' [48]. This builds around the so-called MMPB(GB)SA method, including configurational entropy and free energy of association, but with important additional features to address reliability and reproducibility. Correctly accounting for entropic contributions is essential for reliably predicting binding free energies in cases where the ligands are diverse and/or flexible with many rotatable bonds. The contributions can be incorporated to the calculated free energies using normal mode (NMODE) approach or a variety of other options [68]. We have found a varying number of replicas may be required to achieve a desired level of precision; for many small-molecule–protein systems, 25 replicas are typically required with 4 ns production run for each replica, for ESMACS studies [48,49]. As already noted, the combination of the simulation length and the size of the ensemble provides a trade-off between computational cost and precision. In collaboration with several pharmaceutical companies, we have used the ESMACS approach to investigate drug-like small-molecules bound to therapeutic targets [65–68] and shown that ESMACS is well suited for use in the initial hit-to-lead activities within drug discovery.

(d) Ensemble method for alchemical approaches

The alchemical approach most commonly calculates relative binding free energies between two physical states which are linked by an 'alchemical' path. A series of nonphysical steps are involved in the path. The two physical states can be a protein binding with two ligands, or a ligand binding with wild-type and mutant proteins. Along the alchemical path, some atoms change their chemical identities—appearing, disappearing or alchemically transforming from one to another. Although the alchemical free energy methods are formally exact and general, the possible large uncertainties and expensive computational cost limit the domain of application: they are applicable mainly to estimating small relative free energy changes for structures (drugs or proteins) which involve relatively minor (perturbative) variations. The alchemical method can also be used to calculate absolute binding free energies [73]. It is the equivalent of the relative free energy calculation when one of the ligands involved is replaced by nothing and thus faces even more demanding challenges to achieve convergence.

Free energy calculations using such alchemical methods had rarely been used seriously in drug development projects until recently when Schrödinger Inc. released their 'FEP +' simulation software for relative free energy calculations [95]. With the improved methodology, much of which is proprietary and thus not available for critical assessment, and the use of graphical processing units (GPUs), FEP+ has made a significant impact in the pharmaceutical industry within its domain of applicability [3], although further evaluation is still needed on its accuracy and precision [44,64,73]. From the perspective of this study, however, it is interesting to observe that the methodology advocated is decidedly based on the use of 'one-off' simulations, so that any attempt to provide uncertainty quantification is entirely lacking.

(e) Errors in alchemical calculations

As in many other approaches, an alchemical calculation certainly generates random errors, and very likely contains systematic errors too. As we have stated above, we need to correctly handle the stochastic errors before we can reliably estimate the possible systematic errors. A survey of publications and binding databases shows that the binding affinities are in a range between -6.5 and -15.2 kcal/mol for most of the interesting biomolecular ligands [96].

Molecular simulations aim to predict free energies accurately, with an error approximately 1 kcal/mol, for small molecules to their target proteins. Thermodynamic integration can produce significant differences from individual replica simulations, up to 1.58 kcal/mol and approximately 7 kcal/mol for relative and absolute binding free energies, respectively, for the cases tested using five independent simulations [73]. These simulations, which use the TIES method (see section (f) below), vary only in their initial velocities which are randomly drawn from a Maxwell–Boltzmann distribution. Similar results are obtained from multiple runs of FEP+ calculations, in which up to 3.9 kcal/mol variations have been observed from 30 independent simulations, much larger than the MBAR (multistate Bennett acceptance ratio) errors reported for individual FEP+ calculations [64].

When random errors are handled correctly, it is possible to identify the systematic errors intrinsic to these simulation methods, provided due attention is paid to the way in which errors, both theoretical and experimental, are handled. Systematic errors tend to shift all of the measurements/predictions for the same target systems with the same set-up in the same direction from their real values. Different techniques have been applied in alchemical free energy simulations to enhance sampling in the hope of reducing errors in the calculations. Widely used techniques include accelerated sampling methods, such as replica exchange with solute tempering (REST2) [97], and free energy estimators, such as MBAR [98]. However, our studies paying careful attention to uncertainty quantification show that these techniques offer no guarantee of improving the accuracy of the predictions [44]. Indeed REST2, as used in FEP+, appears to generate a significant systematic underestimation of free energy differences, which degrades monotonically with the duration of simulation [64].

(f) Ensemble alchemical simulations

Alongside ESMACS, we have developed an ensemble-based approach called TIES (thermodynamic integration with enhanced sampling) [74]. TIES employs an ensemble of independent MD simulations in combination to yield accurate and precise free energy predictions. It quantifies and reduces random errors, making the results precise and reproducible. This approach also makes it possible to distinguish systematic errors and to interpret the results correctly.

As one example among many, the application of TIES to protein mutations provides insights underpinning the impact of the gatekeeper mutation of the FGFR-1 kinase on drug efficacy [73]. Using an ensemble-based approach, we were able to quantify the uncertainties in the free energy calculations and to compare the performance of different software and hardware for the calculation of the same free energy changes [64]. Ensemble approaches like TIES provide a reliable and rapid method for uncertainty quantification applied to both relative and absolute binding free energy calculations using alchemical methods [64,73].

6. Materials simulation

Predicting the properties of modern advanced materials typically requires understanding the structure and the dynamical processes on an atomistic level [77,99]. Many large-scale, macroscopic, engineering properties can be modelled using methods taken from continuum mechanics, such as the finite-element method. However, diffusive processes, self-assembly, structural degradation, surface/interface characteristics, and many other quantities of interest to modern materials scientists and engineers are all heavily influenced if not controlled by dynamical processes on the scale of atoms and molecules. This is particularly clear in the case of nanocomposites, in which one has to deal with a polymer matrix in which is embedded a nanomaterial such as a graphene (and its oxide), carbon nanotubes and clays [75,81]. Graphene and other so-called two-dimensional materials have one dimension which is of the order of nanometers, and thus just one or a few atoms thick, yet they impart dramatically enhanced large-scale materials properties in the composites they produce. In such circumstances, it is clear that one must use MD as part of the range of techniques available for studying such complex systems.

MD techniques are uniquely equipped to explore processes that occur on time and length scales of nanometers to microns, and nanoseconds to microseconds. For complex systems, especially those with anisotropic structures on the nanoscale, such as 'soft matter systems', MD has proven useful for predicting the nanoscale structure and material properties. In the development of new materials, the chemical constituents are often among the first known aspects of the system, and the subsequent time necessary for developing useful applications is spent optimizing the fabrication and processing for engineering tests. MD can often help to reduce if not remove many of these practical barriers and assess a material's suitability for a given purpose based only on its atomistic structure. Indeed, there is substantial interest in many areas of materials design in virtual testing using computer simulation to speed up the process from concept to real-world implementation, which currently takes of order 20 years, at a cost of many billions of dollars [100]. The challenge then becomes providing reliable computer-based 'in silico' predictions which reduce the need for expensive and time-consuming experimental work. This puts a premium on providing tight error bars since these furnish a key measure of the confidence with which we can accept modelling results and use them to guide experimental work.

(a) Materials property prediction

By way of example, consider trying to predict a material's stiffness using molecular dynamics (figure 2) [101]. This is one of the most common applications of MD in materials science. A uniaxial stretch of such a system is a fairly trivial simulation to perform; however, it poses several fundamental questions about the certainty we can have in a result of an MD simulation. As discussed above, several systematic uncertainties exist with this technique which is hard to quantify [102], the most glaring example of which is the choice of force field used to represent the material [103]. The often-quoted limitations of MD—such as finite size/time effects or structure generation—are also systematic errors. Using appropriate workflows we can quantify these effects to produce accurate results [104].

MD simulations in the condensed phase are typically performed by imposing periodic boundary conditions in all three spatial directions, which means that we only expect to simulate a comparatively small simulation cell to approximate the bulk properties. The size of this simulation cell has many implications for computational cost but, more importantly, the reliability of the scientific results it furnishes. Finite size effects and fluctuations can be expected to affect the outcome of a simulation.

To measure the Young's modulus (YM) of a material system, the pressure exerted along one axis is sampled before and after (or during) imposition of a small strain. Since the instantaneous pressure of a molecular dynamics simulation can fluctuate by several GPa, it is necessary to average this value over a long sample period to measure the change in pressure due to an applied strain. In a recent study [29], we considered an epoxy resin system, a thermosetting polymer. The investigation quantified the effect of specific MD parameters on the measured Young's modulus, including the system size, starting velocities and polymer generation random seed.

We found that the mean YM of an ensemble of simulations is independent of simulation size but, below a box of size 4 nm, there is a finite size effect which makes the system artificially stiffer (figure 3). This effect only became evident after performing 300 replica simulations at each box size, with mean YM and standard deviation of 3.4 ± 1.9 GPa. The smallest box size gave a distribution of YMs with a significant skewness of -0.8; as the box size increased the skewness tended to zero. The distribution is effectively Gaussian because there are no long-range interactions present. The analysis was greatly facilitated by the use of the EasyVVUQ software [29]. From this, it should be clear that one single measurement, at the low strains imposed here, is wholly inadequate to measure this property reliably.

Another benefit of running ensemble simulations is that one can perform sensitivity analysis and thereby learn about the variance in properties arising due to different input variables and parameters. By applying the law of total variance, our study showed that the expected variance due to the polymer network generation was equal to that due to the starting velocities of the atoms
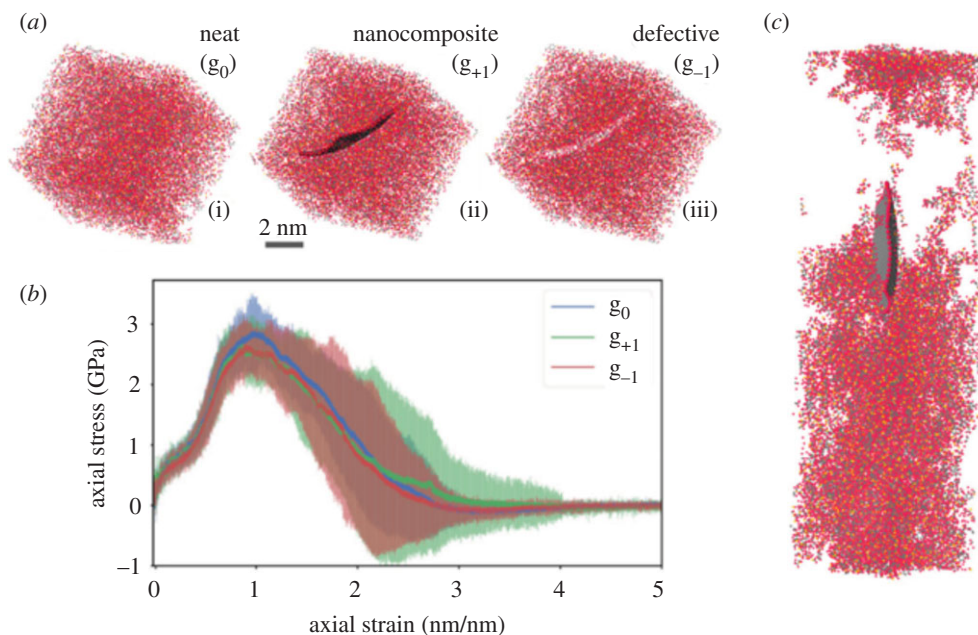
Figure 2. Measuring the toughness of different materials with a reactive forcefield. (*a*) Three structures: (i) neat epoxy polymer; (ii) epoxy-graphene nanocomposite and (iii) epoxy polymer with a defect are strained uniaxially. (*b*) The stress–strain curves are shown in the plot; lines indicate the average of six replica simulations while the shaded regions correspond to the standard deviations at each strain. While each replica varies, the ensemble average shows the three materials behave similarly. (*c*) Displays a snapshot from sample (ii) at the point of fracture. (Online version in colour.)
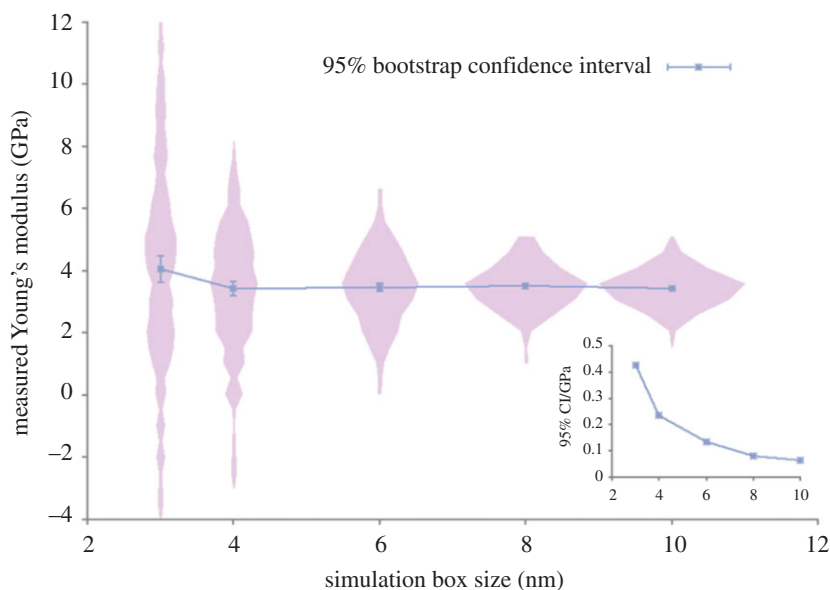

Figure 3. Young's modulus of an epoxy resin measured with different simulation sizes. Each point is the average of 300 simulations, which make up the pink histograms for each box size. The 95% bootstrap confidence interval for increasing ensemble size is shown in the inset plot at the bottom right. (Online version in colour.)

in the simulation. In other words, the exact connectivity of monomer units was inconsequential, provided that the same cross-link density was achieved. However, one single simulation of such a polymer is insufficient; the aforementioned ensemble of 300 replicas was required in order to be 95% confident of the size effect seen above.

These approaches are broadly applicable. Running ensembles that give statistically significant results not only allows us to efficiently sample more phase space, therefore increasing the accuracy in the result, but provides much more information that single simulations would not be able to tell us.

Ensemble techniques are necessary for exploring phase space whenever molecular dynamics simulations are used. Alfè *et al.* explored the kinetics of phase transitions in superheated metals caused by nucleation processes [105,106]. The time taken for a superheated metal to melt in this manner is generally unknown: by simulating an ensemble of 350 replicas, differing only in the starting velocities of the atoms concerned, they were able to make accurate predictions of the system's kinetics. They found that nucleation rates are highly dependent on the details of nanoscale behaviour. Here too, finite size effects were isolated and corrected for in reporting reliable results.

(b) Generating structures

It is often forgotten that the starting structure can itself be a major source of uncertainty in MD simulations; it is as true for materials as it is in biomedical simulation. The most straightforward way of achieving variation in an ensemble is to use different random numbers to seed the initial atomic velocities; however, generating different starting configurations should also be considered as we did when performing an MD study of the Nobel prize-winning discovery of graphene by peeling off atomic layers from graphite [80]. The process of building the initial structure and coordinates of a system is not trivial and can often be the most time-consuming step in such research work: it may involve non-trivial polymer chain building for synthetic and biological macromolecular structures, the details about the microstructure of nanocomposites, and so on. The initial state of an MD simulation is inevitably artificial and therefore not itself a representation of the system of interest. Initial states must be built such that, for example, if one wishes to study the equilibrium state, it will not take an inordinately long time to reach that state by MD simulation. The starting structure must be sampled sufficiently and carefully assessed to check that unphysical starting conditions do not influence the production stage of a simulation.

Polymer systems are manifestly difficult to build while diffusion in high molecular weight polymers can be extremely slow, so entanglements and anisotropic structures must be built carefully as good starting points. Numerous techniques exist to do this. Diffusive processes are so slow that 'sampling the phase space' with one long simulation would be extremely expensive using normal molecular simulation approaches and inaccurate into the bargain; instead several structure generations are essential to be confident in a result.

Graphene oxide poses a different problem as its structure is that of an amorphous crystal [107]. Oxygen-containing groups are present across the surface of graphene with some random distribution; in the presence of other materials, the precise distribution of these groups may influence their interaction. In the case of a graphene oxide dissolved in a polymer melt, it is not sufficient to build one structure and generate an ensemble with different initial velocities; instead one must generate several graphene oxide structures to understand the system.

(c) Forcefield critical properties

The errors caused by inaccurate force fields can be completely catastrophic for the physics under investigation. A stable inaccuracy in a force field may give binding energy within some error of the true value, but a more significant inaccuracy may result in divergent dynamical and/or structural regimes [108]. For example, when predicting the structure of crystals, a flawed force field may never produce certain crystal arrangements. Lennard-Jones forcefields are known to

underestimate the friction between layered materials [109]. Sinclair *et al.* [75] found that the spherical symmetry of these potentials is too gross an approximation while simulating graphene bilayers so a new forcefield was developed. In the experiment, graphene flakes are observed to show superlubric behaviour when propelled across a graphitic surface. Propelling flakes in this way is a chaotic process (in the technical sense that it is highly dependent on the initial conditions). In order to achieve an acceptable error which was comparable with experimental data on frictional properties, ensembles of 40 replicas were required. The distance travelled by a flake seems to follow a lognormal distribution, with a Fisher–Pearson coefficient of skewness of 1.4.

7. Generating actionable predictions

In order for the predictions from computational science to inform costly and consequential decisions for real-world problems, it is vital that they are accurate, reproducible and accompanied by uncertainty quantification. Speed too is of critical concern, in order for predictions to be actionable—that enables decision makers to take appropriate actions in a certain period of time. Aerospace manufacturers are keen on the concept of virtual certification in order to reduce the time to market, along the way from concept to implementation. Approaches and tools have been proposed to perform uncertainty analyses in real-world practice [14]. Emulator based methods [14] and the test harness [110] are used to analyse sensitivity and uncertainty, to evaluate scientific software and to calibrate complex computer simulators. Practical recommendations have been made on validation and reproducibility in the field of scientific research in general [111] including molecular dynamics [5,39]. Uncorrelated data need to be collected in sufficient quantities; autocorrelation analyses can be used to better understand if a time series represents an equilibrated system [111], but caution must be exercised as longer autocorrelation times may be uncovered if other relevant free energy minima are revealed [45].

In computer-aided drug discovery, different levels of uncertainties are needed in various stages, including searching for potential small-molecule hits, identifying promising leads and optimizing leads for further evaluation. Concentrating on the objects of uncertainty can give us an appreciation of how decisions can be made based on uncertainty quantification and cost-benefit analyses. The Accelerating Therapeutics for Opportunities in Medicine (ATOM) project (https://atomscience.org/), funded by the Department of Energy in the USA, is predicated on the aim of being able to go from concept to clinical trials for new drugs within 12 months. This is a very large and highly ambitious project involving many partners, from academia, industry and US national laboratories. Being able to rely on the results of *in silico* predictions will undoubtedly turn these industries on their heads, by eliminating large amounts of laboratory-based testing. Knowledge of the uncertainties attaching to decisions is critical here.

In the field of medicine, where patient safety is at stake, it is extremely important to have an uncertainty quantification discipline for risk management [112]. We are active in the context of clinical decision making for personalized medicine and have discussed these requirements for many years (see e.g. Groen *et al.* [113] and Manos *et al.* [114]). Reproducibility and uncertainty (indeed, more completely, VVUQ) are of central concern for the predictions to be deemed actionable: for the uptake of *in silico* approaches in medical and clinical contexts, regulatory authorities will demand procedures which are fully certified in this sense. This can be done through careful control of the uncertainties in the calculations employed, which rest heavily on molecular dynamics computation.

Generating actionable predictions requires a high level of automation which can be achieved through a powerful combination of software and hardware, making calculations immediately scalable for industrial and clinical applications. To make predictions which are fast enough for actionable decision making, urgent priority must be given to such calculations on high-performance computers. Our requirements for on demand access to large-scale computing resources for such purposes are well known and have informed numerous initiatives across the world to provide appropriate access mechanisms for medical and clinical research. This is part of

the core business of the Computational Biomedicine Centre of Excellence (www.compbiomed.eu) which has innovation and sustainability at its heart. As a matter of fact, the EasyVVUQ software which we have introduced as an open-source toolkit, and as part of the VECMA Toolkit [30], is based on its original applications to MD as discussed here. It is now being applied widely in many fields of concern to this theme issue of *Phil Trans A*, from fluid dynamics to climate prediction and fusion energy research. As such, it can be reasonably considered that the study of uncertainty quantification in MD is now beginning to have an impact within the wider general field of VVUQ.

8. Conclusion

Ensemble molecular dynamics simulation provides us with a powerful methodology that enables us to connect ergodic theory and uncertainty quantification, and to obtain reproducible results from simulations in a systematic and theoretically well-grounded manner. As evidenced in the case of ligand–protein binding free energy predictions, ensemble simulation-based approaches yield statistically robust, precise and reproducible, hence reliable results. Using ensemble methods, the errors in predictions can be systematically controlled, amenable to further reduction by increasing the number of replicas in an ensemble and in propitious circumstances too by extending the length of such simulations. Ensemble approaches are scalable: for example, they permit hundreds to thousands of binding affinities to be calculated per day, depending on the computing resources available. Computing capabilities are set to increase as the exascale era heaves in to view. In the near future, rapid, accurate and reliable predictions of materials properties may emerge that can be exploited in the aerospace and automotive industries; free energy prediction at high throughput will assist physicians in clinical decision making and medicinal chemists in directing compound synthesis in a routine manner. In sectors such as these, virtual certification and regulatory approval for the use of *in silico* methods will depend critically on the application of rigorous uncertainty quantification along the lines we have described.

References

1. Alder BJ, Wainwright TE. 1957 Phase transition for a hard sphere system. *J. Chem. Phys.* **27**, 1208–1209. (doi:10.1063/1.1743957)
2. Hu Z, Mahadevan S. 2017 Uncertainty quantification and management in additive manufacturing: current status, needs, and opportunities. *Int. J. Adv. Manuf. Technol.* **93**, 2855–2874. (doi:10.1007/s00170-017-0703-5)
3. Sherborne B *et al.* 2016 Collaborating to improve the use of free-energy and other quantitative methods in drug discovery. *J. Comput. Aided Mol. Des.* **30**, 1139–1141. (doi:10.1007/s10822-016-9996-y)

4. Coveney PV, Highfield RR. 2021 When we can trust computers (and When We Can't). *Phil. Trans. R. Soc. A* **379**, 20200067. (doi:10.1098/rsta.2020.0067)

5. Patrone PN, Dienstfrey A. 2018 Uncertainty quantification for molecular dynamics. In *Reviews in computational chemistry*, pp. 115–169. Hoboken, NJ: John Wiley & Sons.

6. Molteni F, Buizza R, Palmer TN, Petroliagis T. 1996 The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119. (doi:10.1002/qj.49712252905)

7. Suleimenova D, Groen D. 2020 How policy decisions affect refugee journeys in south sudan: a study using automated ensemble simulations. *J. Artif. Soc. Soc. Simul.* **23**, 2. (doi:10.18564/jasss.4193)

8. Wan S, Bhati AP, Zasada SJ, Coveney PV. 2020 Rapid, accurate, precise and reproducible ligand-protein binding free energy prediction. *Interface Focus* **10**, 20200007. (doi:10.1098/rsfs.2020.0007)

9. Wright DW, Wan S, Shublaq N, Zasada SJ, Coveney PV. 2012 From base pair to bedside: molecular simulation and the translation of genomics to personalized medicine. *Wiley Interdiscip Rev. Syst. Biol. Med.* **4**, 585–598. (doi:10.1002/wsbm.1186)

10. Fowler PW. 2021 How quickly can we predict trimethoprim resistance using alchemical free energy methods? *Interface Focus* **10**, 20190141. (doi:10.1098/rsfs.2019.0141)

11. Adam D. 2020 Special report: the simulations driving the world's response to COVID-19. *Nature* **580**, 316–318. (doi:10.1038/d41586-020-01003-6)

12. Ferguson NM *et al.* 2020 Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial College London (published online). (doi:10.25561/77482)

13. Edeling W *et al.* 2020 Model uncertainty and decision making: predicting the impact of COVID-19 using the CovidSim epidemiology code. See https://www.researchsquare.com/article/rs-82122/v3 (accessed 29 December 2020).

14. Dienstfrey AM, Boisvert RF (eds). 2012 *Uncertainty quantification in scientific computing*, vol. 377. Berlin, Germany: IFIP Advances in Information and Communication Technology.

15. Mortensen JJ, Kaasbjerg K, Frederiksen SL, Nørskov JK, Sethna JP, Jacobsen KW. 2005 Bayesian error estimation in density-functional theory. *Phys. Rev. Lett.* **95**, 216401. (doi:10.1103/PhysRevLett.95.216401)

16. Simm GN, Reiher M. 2016 Systematic error estimation for chemical reaction energies. *J. Chem. Theory Comput.* **12**, 2762–2773. (doi:10.1021/acs.jctc.6b00318)

17. Walker E, Ammal SC, Terejanu GA, Heyden A. 2016 Uncertainty quantification framework applied to the water–gas shift reaction over Pt-based catalysts. *J. Phys. Chem. C.* **120**, 10 328–10 339. (doi:10.1021/acs.jpcc.6b01348)

18. Proppe J, Husch T, Simm GN, Reiher M. 2016 Uncertainty quantification for quantum chemical models of complex reaction networks. *Faraday Discuss.* **195**, 497–520. (doi:10.1039/C6FD00144K)

19. Pernot P. 2017 The parameter uncertainty inflation fallacy. *J. Chem. Phys.* **147**, 104102. (doi:10.1063/1.4994654)

20. Proppe J, Reiher M. 2017 Reliable estimation of prediction uncertainty for physicochemical property models. *J. Chem. Theory Comput.* **13**, 3297–3317. (doi:10.1021/acs.jctc.7b00235)

21. Pernot P, Savin A. 2020 Probabilistic performance estimators for computational chemistry methods: systematic improvement probability and ranking probability matrix. *I. Theory J. Chem. Phys.* **152**, 164108. (doi:10.1063/5.0006202)

22. Scalia G, Grambow CA, Pernici B, Li Y-P, Green WH. 2020 Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *J. Chem. Inf. Model.* **60**, 2697–2717. (doi:10.1021/acs.jcim.9b00975)

23. Tran K, Neiswanger W, Yoon J, Zhang Q, Xing E, Ulissi ZW. 2020 Methods for comparing uncertainty quantifications for material property predictions. *Mach. Learn Sci. Technol.* **1**, 025006. (doi:10.1088/2632-2153/ab7e1a)

24. Simm GN, Reiher M. 2018 Error-controlled exploration of chemical reaction networks with Gaussian processes. *J. Chem. Theory Comput.* **14**, 5238–5248. (doi:10.1021/acs.jctc.8b00504)

25. Proppe J, Gugler S, Reiher M. 2019 Gaussian process-based refinement of dispersion corrections. *J. Chem. Theory Comput.* **15**, 6046–6060. (doi:10.1021/acs.jctc.9b00627)

26. Jorgensen WL. 2004 The many roles of computation in drug discovery. *Science* **303**, 1813–1818. (doi:10.1126/science.1096361)
27. Cournia Z, Allen B, Sherman W. 2017 Relative binding free energy calculations in drug discovery: recent advances and practical considerations. *J. Chem. Inf. Model.* **57**, 2911–2937. (doi:10.1021/acs.jcim.7b00564)
28. The Alan Turing Institute. 2020 Reliability and reproducibility in computational science: Implementing verification, validation and uncertainty quantification in silico. See https://www.turing.ac.uk/events/reliability-and-reproducibility-computational-science
29. Wright DW *et al.* 2020 Building confidence in simulation: applications of EasyVVUQ. *Adv. Theory Simul*. 3, 1900246. (doi:10.1002/adts.201900246)
30. Groen D *et al.* 2019 Introducing VECMAtk - Verification, Validation and Uncertainty Quantification for Multiscale and HPC Simulations, pp. 479–492.
31. Leimkuhler B, Matthews C. 2015 *Molecular dynamics*, vol. 39. Cham, Switzerland: Springer International Publishing.
32. Van Noorden R. 2013 Modellers react to chemistry award. *Nature* **502**, 280. (doi:10.1038/502280a)
33. Mobley DL, Wymer KL, Lim NM, Guthrie JP. 2014 Blind prediction of solvation free energies from the SAMPL4 challenge. *J. Comput. Aided Mol. Des.* **28**, 135–150. (doi:10.1007/s10822-014-9718-2)
34. Bannan CC, Burley KH, Chiu M, Shirts MR, Gilson MK, Mobley DL. 2016 Blind prediction of cyclohexane–water distribution coefficients from the SAMPL5 challenge. *J. Comput. Aided Mol. Des.* **30**, 927–944. (doi:10.1007/s10822-016-9954-8)
35. Rizzi A *et al.* 2018 Overview of the SAMPL6 host–guest binding affinity prediction challenge. *J. Comput. Aided Mol. Des.* **32**, 937–963. (doi:10.1007/s10822-018-0170-6)
36. Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, Shaw DE. 2012 Systematic validation of protein force fields against experimental data. Muller DJ, editor. *PLoS ONE* **7**, e32131. (doi:10.1371/journal.pone.0032131)
37. Best RB, Buchete N-V, Hummer G. 2008 Are current molecular dynamics force fields too helical? *Biophys. J.* **95**, L7–L9. (doi:10.1529/biophysj.108.132696)
38. Lopes PEM, Guvench O, MacKerell AD. 2015 *Current status of protein force fields for molecular dynamics simulations*, pp. 47–71. New York, NY: Humana Press.
39. van Gunsteren WF *et al.* 2018 Validation of molecular simulation: an overview of issues. *Angew Chemie Int. Ed.* **57**, 884–902. (doi:10.1002/anie.201702945)
40. Hansen N, van Gunsteren WF. 2014 Practical aspects of free-energy calculations: a review. *J. Chem. Theory Comput.* **10**, 2632–2647. (doi:10.1021/ct500161f)
41. Boghosian BM, Coveney PV, Wang H. 2019 A new pathology in the simulation of chaotic dynamical systems on digital computers. *Adv. Theory Simul.* **2**, 1900125. (doi:10.1002/adts.201900125)
42. Coveney PV, Wan S. 2016 On the calculation of equilibrium thermodynamic properties from molecular dynamics. *Phys. Chem. Chem. Phys.* **18**, 30 236–30 240. (doi:10.1039/C6CP02349E)
43. Leitgeb M, Schröder C, Boresch S. 2005 Alchemical free energy calculations and multiple conformational substates. *J. Chem. Phys.* **122**, 084109. (doi:10.1063/1.1850900)
44. Bhati AP, Wan S, Coveney PV. 2019 Ensemble-based replica exchange alchemical free energy methods: the effect of protein mutations on inhibitor binding. *J. Chem. Theory Comput.* **15**, 1265–1277. (doi:10.1021/acs.jctc.8b01118)
45. Gapsys V, de Groot BL. 2020 On the importance of statistics in molecular simulations for thermodynamics, kinetics and simulation box size. *Elife* **9**, e57589. (doi:10.7554/eLife.57589)
46. Sadiq SK, Wright DW, Kenway OA, Coveney PV. 2010 Accurate ensemble molecular dynamics binding free energy ranking of multidrug-resistant HIV-1 proteases. *J. Chem. Inf. Model.* **50**, 890–905. (doi:10.1021/ci100007w)
47. Wan S, Coveney PV. 2011 Rapid and accurate ranking of binding affinities of epidermal growth factor receptor sequences with selected lung cancer drugs. *J. R Soc. Interface.* **8**, 1114–1127. (doi:10.1098/rsif.2010.0609)
48. Wan S, Knapp B, Wright DW, Deane CM, Coveney PV. 2015 Rapid, precise, and reproducible prediction of peptide-MHC binding affinities from molecular dynamics that correlate well with experiment. *J. Chem. Theory Comput.* **11**, 3346–3356. (doi:10.1021/acs.jctc.5b00179)

49. Wright DW, Hall BA, Kenway OA, Jha S, Coveney PV. 2014 Computing clinically relevant binding free energies of HIV-1 protease inhibitors. *J. Chem. Theory Comput.* **10**, 1228–1241. (doi:10.1021/ct4007037)

50. Succi S, Coveney PV. 2019 Big data: the end of the scientific method? *Phil. Trans. R. Soc. A. Math. Phys. Eng. Sci.* **377**, 20180145. (doi:10.1098/rsta.2018.0145)

51. El Hage K, Hédin F, Gupta PK, Meuwly M, Karplus M. 2018 Valid molecular dynamics simulations of human hemoglobin require a surprisingly large box size. *Elife* **7**, e35560. (doi:10.7554/eLife.35560)

52. Gapsys V, de Groot BL. 2019 Comment on 'Valid molecular dynamics simulations of human hemoglobin require a surprisingly large box size'. *Elife* 8, e44718. (doi:10.7554/eLife.44718)

53. El Hage K, Hédin F, Gupta PK, Meuwly M, Karplus M. 2019 Response to comment on 'Valid molecular dynamics simulations of human hemoglobin require a surprisingly large box size'. *Elife* 8, e45318. (doi:10.7554/eLife.45318)

54. Gapsys V, de Groot BL. 2019 Comment on 'Response to comment on 'Valid molecular dynamics simulations of human hemoglobin require a surprisingly large box size.' *bioRxiv.* 2019, 812966. (doi:10.1101/812966)

55. Pezzella M, El Hage K, Niesen MJM, Shin S, Willard AP, Meuwly M, Karplus M. 2020 Water dynamics around proteins: T- and R-states of hemoglobin and melittin. *J. Phys. Chem. B.* **124**, 6540–6554. (doi:10.1021/acs.jpcb.0c04320)

56. Caves LSD, Evanseck JD, Karplus M. 1998 Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci.* **7**, 649–666. (doi:10.1002/pro.5560070314)

57. Elofsson A, Nilsson L. 1993 How consistent are molecular dynamics simulations? *J. Mol. Biol.* **233**, 766–780. (doi:10.1006/jmbi.1993.1551)

58. Laio A, Parrinello M. 2002 Escaping free-energy minima. *Proc. Natl Acad. Sci. USA* **99**, 12 562–12 566. (doi:10.1073/pnas.202427399)

59. Hamelberg D, Mongan J, McCammon JA. 2004 Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **120**, 11 919–11 929. (doi:10.1063/1.1755656)

60. Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC. 2007 Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.* **126**, 155101. (doi:10.1063/1.2714538)

61. Zheng L, Chen M, Yang W. 2009 Simultaneous escaping of explicit and hidden free energy barriers: application of the orthogonal space random walk strategy in generalized ensemble based conformational sampling. *J. Chem. Phys.* **130**, 234105. (doi:10.1063/1.3153841)

62. Wu X, Brooks BR. 2003 Self-guided Langevin dynamics simulation method. *Chem. Phys. Lett.* **381**, 512–518. (doi:10.1016/j.cplett.2003.10.013)

63. Fukunishi H, Watanabe O, Takada S. 2002 On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: application to protein structure prediction. *J. Chem. Phys.* **116**, 9058–9067. (doi:10.1063/1.1472510)

64. Wan S, Tresadern G, Pérez-Benito L, Vlijmen H, Coveney PV. 2020 Accuracy and precision of alchemical relative free energy predictions with and without replica-exchange. *Adv. Theory Simul* **3**, 1900195. (doi:10.1002/adts.201900195)

65. Wan S, Bhati AP, Skerratt S, Omoto K, Shanmugasundaram V, Bagal SK, Coveney PV. 2017 Evaluation and characterization of Trk kinase inhibitors for the treatment of pain: reliable binding affinity predictions from theory and computation. *J. Chem. Inf. Model.* **57**, 897–909. (doi:10.1021/acs.jcim.6b00780)

66. Wan S, Bhati AP, Zasada SJ, Wall I, Green D, Bamborough P, Coveney PV. 2017 Rapid and reliable binding affinity prediction of bromodomain inhibitors: a computational study. *J. Chem. Theory Comput.* **13**, 784–795. (doi:10.1021/acs.jctc.6b00794)

67. Wright DW, Husseini F, Wan S, Meyer C, van Vlijmen H, Tresadern G, Coveney PV. 2019 Application of the ESMACS binding free energy protocol to a multi-binding site lactate dehydogenase a ligand dataset. *Adv. Theory Simul.* 3, 1900194. (doi:10.1002/adts.201900194)

68. Wright DW, Wan S, Meyer C, van Vlijmen H, Tresadern G, Coveney PV. 2019 Application of ESMACS binding free energy protocols to diverse datasets: bromodomain-containing protein 4. *Sci. Rep.* **9**, 1–15. (doi:10.1038/s41598-019-41758-1)

69. Bergonzo C, Henriksen NM, Roe DR, Swails JM, Roitberg AE, Cheatham TE. 2014 Multidimensional replica exchange molecular dynamics yields a converged ensemble of an RNA tetranucleotide. *J. Chem. Theory Comput.* **10**, 492–499. (doi:10.1021/ct400862k)

70. Lyman E, Zuckerman DM. 2007 On the structural convergence of biomolecular simulations by determination of the effective sample size. *J. Phys. Chem. B.* **111**, 12 876–12 882. (doi:10.1021/jp073061t)

71. Fowler PW, Cole K, Gordon NC, Kearns AM, Llewelyn MJ, Peto TEA, Crook DW, Walker AS. 2018 Robust prediction of resistance to trimethoprim in Staphylococcus aureus. *Cell Chem. Biol.* **25**, 339–349. (doi:10.1016/j.chembiol.2017.12.009)

72. Knapp B, Ospina L, Deane CM. 2018 Avoiding false positive conclusions in molecular simulation: the importance of replicas. *J. Chem. Theory Comput.* **14**, 6127–6138. (doi:10.1021/acs.jctc.8b00391)

73. Bhati AP, Wan S, Hu Y, Sherborne B, Coveney PV. 2018 Uncertainty quantification in alchemical free energy methods. *J. Chem. Theory Comput.* **14**, 2867–2880. (doi:10.1021/acs.jctc.7b01143)

74. Bhati AP, Wan S, Wright DW, Coveney PV. 2017 Rapid, accurate, precise, and reliable relative free energy prediction using ensemble based thermodynamic integration. *J. Chem. Theory Comput.* **13**, 210–222. (doi:10.1021/acs.jctc.6b00979)

75. Sinclair RC, Suter JL, Coveney PV. 2018 Graphene–graphene interactions: friction, superlubricity, and exfoliation. *Adv. Mater.* **30**, 1705791. (doi:10.1002/adma.201705791)

76. Potterton A, Husseini FS, Southey MWY, Bodkin MJ, Heifetz A, Coveney PV *et al.* 2019 Ensemble-based steered molecular dynamics predicts relative residence time of A_{2A} receptor binders. *J. Chem. Theory Comput.* **15**, 3316–3330. (doi:10.1021/acs.jctc.8b01270)

77. Vassaux M, Sinclair RC, Richardson RA, Suter JL, Coveney PV. 2019 Toward high fidelity materials property prediction from multiscale modeling and simulation. *Adv. Theory Simul.* **1900122**, 1–18. (doi:10.1002/adts.201900122)

78. Amaro RE, Baudry J, Chodera J, Demir Ö, McCammon JA, Miao Y, Smith JC. 2018 Ensemble docking in drug discovery. *Biophys. J.* **114**, 2271–2278. (doi:10.1016/j.bpj.2018.02.038)

79. Saadi AA *et al.* 2020 IMPECCABLE: Integrated Modeling PipelinE for COVID Cure by Assessing Better LEads. See http://arxiv.org/abs/2010.06574.

80. Sinclair RC, Suter JL, Coveney PV. 2019 Micromechanical exfoliation of graphene on the atomistic scale. *Phys. Chem. Chem. Phys.* **21**, 5716–5722. (doi:10.1039/C8CP07796G)

81. Suter JL, Sinclair RC, Coveney PV. 2020 Principles governing control of aggregation and dispersion of graphene and graphene oxide in polymer melts. *Adv. Mater.* **28**, 2003213. (doi:10.1002/adma.202003213)

82. Ahmad K. 2020 Ensemble-Based Coarse-Grained Molecular Dynamics Simulations of Multifunctional DNA Nanopores, PhD Thesis, University College London.

83. Efron B, Tibshirani RJ. 1994 *An introduction to the bootstrap*. Oxford, UK: Chapman and Hall/CRC Press.

84. Sadiq SK, Wright D, Watson SJ, Zasada SJ, Stoica I, Coveney PV. 2008 Automated molecular simulation based binding affinity calculator for ligand-bound HIV-1 proteases. *J. Chem. Inf. Model.* **48**, 1909–1919. (doi:10.1021/ci8000937)

85. VECMA Toolkit. 2020 See https://www.vecma-toolkit.eu/.

86. Richardson RA, Wright DW, Edeling W, Jancauskas V, Lakhlili J, Coveney PV. 2020 EasyVVUQ: a library for verification, validation and uncertainty quantification in high performance computing. *J. Open Res. Softw.* **8**, 11. (doi:10.5334/jors.303)

87. Hummer G, Pratt LR, García AE. 1996 Free energy of ionic hydration. *J. Phys. Chem.* **100**, 1206–1215. (doi:10.1021/jp951011v)

88. Shirts MR, Pande VS. 2005 Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. *J. Chem. Phys.* **122**, 144107. (doi:10.1063/1.1873592)

89. König G, Brooks BR, Thiel W, York DM. 2018 On the convergence of multi-scale free energy simulations. *Mol. Simul.* **44**, 1062–1081. (doi:10.1080/08927022.2018.1475741)

90. de Ruiter A, Oostenbrink C. 2020 Advances in the calculation of binding free energies. *Curr. Opin Struct. Biol.* **61**, 207–212. (doi:10.1016/j.sbi.2020.01.016)

91. Pérez A, Martínez-Rosell G, De Fabritiis G. 2018 Simulations meet machine learning in structural biology. *Curr. Opin Struct. Biol.* **49**, 139–144. (doi:10.1016/j.sbi.2018.02.004)

92. Åqvist J, Luzhkov VB, Brandsdal BO. 2002 Ligand binding affinities from MD simulations. *Acc. Chem. Res.* **35**, 358–365. (doi:10.1021/ar010014p)

93. Genheden S, Ryde U. 2009 How to obtain statistically converged MM/GBSA results. *J. Comput. Chem.* **31**, 837–846. (doi:10.1002/jcc.21366)

94. Gumbart JC, Roux B, Chipot C. 2013 Standard binding free energies from computer simulations: what is the best strategy? *J. Chem. Theory Comput.* **9**, 794–802. (doi:10.1021/ct3008099)

95. Wang L *et al.* 2015 Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **137**, 2695–2703. (doi:10.1021/ja512751q)

96. Chodera JD, Mobley DL. 2013 Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. *Annu. Rev. Biophys.* **42**, 121–142. (doi:10.1146/annurev-biophys-083012-130318)

97. Wang L, Friesner RA, Berne BJ. 2011 Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). *J. Phys. Chem. B.* **115**, 9431–9438. (doi:10.1021/jp204407d)

98. Shirts MR, Chodera JD. 2008 Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **129**, 124105. (doi:10.1063/1.2978177)

99. Bernardo LFA, Amaro APBM, Pinto DG, Lopes SMR. 2016 Modeling and simulation techniques for polymer nanoparticle composites – a review. *Comput. Mater. Sci.* **118**, 32–46. (doi:10.1016/j.commatsci.2016.02.025)

100. Salter AJ, Martin BR. 2001 The economic benefits of publicly funded basic research: a critical review. *Res. Policy* **30**, 509–532. (doi:10.1016/S0048-7333(00)00091-3)

101. Vassaux M, Sinclair RC, Richardson RA, Suter JL, Coveney PV. 2019 The role of graphene in enhancing the material properties of thermosetting polymers. *Adv. Theory Simul.* **2**, 1800168. (doi:10.1002/adts.201800168)

102. Clavier G, Desbiens N, Bourasseau E, Lachet V, Brusselle-Dupend N, Rousseau B. 2017 Computation of elastic constants of solids using molecular simulation: comparison of constant volume and constant pressure ensemble methods. *Mol. Simul.* **43**, 1413–1422. (doi:10.1080/08927022.2017.1313418)

103. Odegard G, Clancy T, Gates T. 2005 Prediction of mechanical properties of polymers with various force fields. In *46th AIAAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference*. Reston, Viriginal: American Institute of Aeronautics and Astronautics.

104. Reeve ST, Strachan A. 2019 Functional uncertainty quantification for isobaric molecular dynamics simulations and defect formation energies. *Model Simul. Mater. Sci. Eng.* **27**, 044002. (doi:10.1088/1361-651X/ab16fa)

105. Alfè D, Cazorla C, Gillan MJ. 2011 The kinetics of homogeneous melting beyond the limit of superheating. *J. Chem. Phys.* **135**, 024102. (doi:10.1063/1.3605601)

106. Davies CJ, Pozzo M, Alfè D. 2019 Assessing the inner core nucleation paradox with atomic-scale simulations. *Earth Planet. Sci. Lett.* **507**, 1–9. (doi:10.1016/j.epsl.2018.11.019)

107. Sinclair RC, Coveney PV. 2019 Modeling nanostructure in graphene oxide: inhomogeneity and the percolation threshold. *J. Chem. Inf. Model.* **59**, 2741–2745. (doi:10.1021/acs.jcim.9b00114)

108. Ragasa EJ, O'Brien CJ, Hennig RG, Foiles SM, Phillpot SR. 2019 Multi-objective optimization of interatomic potentials with application to MgO. *Model Simul. Mater. Sci. Eng.* **27**, 074007. (doi:10.1088/1361-651X/ab28d9)

109. Jiang J-W, Park HS. 2015 A Gaussian treatment for the friction issue of Lennard-Jones potential in layered materials: application to friction between graphene, MoS_2, and black phosphorus. *J. Appl. Phys.* **117**, 124304. (doi:10.1063/1.4916538)

110. Smith BT. 2012 Measuring uncertainty in scientific computation using numerica 21's test harness. In *Uncertainty Quantification in Scientific Computing*, pp. 165–179.

111. Grossfield A, Patrone PN, Roe DR, Schultz AJ, Siderius D, Zuckerman DM. 2019 Best practices for quantification of uncertainty and sampling quality in molecular simulations [Article v1.0]. *Living J. Comput. Mol. Sci.* **1**, 5067. (doi:10.33011/livecoms.1.1.5067)

112. Begoli E, Bhattacharya T, Kusnezov D. 2019 The need for uncertainty quantification in machine-assisted medical decision making. *Nat. Mach. Intell.* **1**, 20–23. (doi:10.1038/s42256-018-0004-1)

113. Groen D *et al.* 2018 Validation of patient-specific cerebral blood flow simulation using transcranial doppler measurements. *Front. Physiol.* **9**, 1–13. (doi:10.3389/fphys.2018.00721)

114. Manos S, Zasada S, Coveney PV. 2008 Life or death decision-making: the medical case for large-scale, on-demand grid computing. *CTWatch Q.* **4**, 1.

23

royalsocietypublishing.org/journal/rsta *Phil. Trans. R. Soc. A* **379**: 20200082