INTERFACE FOCUS

royalsocietypublishing.org/journal/rsfs

Review

Cite this article: Wan S, Bhati AP, Zasada SJ, Coveney PV. 2020 Rapid, accurate, precise and reproducible ligand–protein binding free energy prediction. *Interface Focus* **10**: 20200007.
http://dx.doi.org/10.1098/rsfs.2020.0007

Accepted: 11 September 2020

One contribution of 9 to a theme issue 'Computational biomedicine. Part I: molecular medicine'.

Subject Areas:
computational biology, bioinformatics, medical physics

Keywords:
binding free energy, ensemble simulation, reproducibility, molecular dynamics

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

THE ROYAL SOCIETY PUBLISHING

Rapid, accurate, precise and reproducible ligand–protein binding free energy prediction

Shunzhou Wan[1], Agastya P. Bhati[1], Stefan J. Zasada[1] and Peter V. Coveney[1,2]

[1]Centre for Computational Science, Department of Chemistry, University College London, London WC1H 0AJ, UK
[2]Computational Science Laboratory, Institute for Informatics, Faculty of Science, University of Amsterdam, 1098XH Amsterdam, The Netherlands

SW, 0000-0001-7192-1999; APB, 0000-0003-4539-4819; SJZ, 0000-0003-4643-4982; PVC, 0000-0002-8787-7256

A central quantity of interest in molecular biology and medicine is the free energy of binding of a molecule to a target biomacromolecule. Until recently, the accurate prediction of binding affinity had been widely regarded as out of reach of theoretical methods owing to the lack of reproducibility of the available methods, not to mention their complexity, computational cost and time-consuming procedures. The lack of reproducibility stems primarily from the chaotic nature of classical molecular dynamics (MD) and the associated extreme sensitivity of trajectories to their initial conditions. Here, we review computational approaches for both relative and absolute binding free energy calculations, and illustrate their application to a diverse set of ligands bound to a range of proteins with immediate relevance in a number of medical domains. We focus on ensemble-based methods which are essential in order to compute statistically robust results, including two we have recently developed, namely thermodynamic integration with enhanced sampling and enhanced sampling of MD with an approximation of continuum solvent. Together, these form a set of rapid, accurate, precise and reproducible free energy methods. They can be used in real-world problems such as hit-to-lead and lead optimization stages in drug discovery, and in personalized medicine. These applications show that individual binding affinities equipped with uncertainty quantification may be computed in a few hours on a massive scale given access to suitable high-end computing resources and workflow automation. A high level of accuracy can be achieved using these approaches.

1. Introduction

The use of computer models and simulations to understand natural systems is now widespread, encompassing many diverse disciplines in academia as well as industry. One of the major advantages of computational modelling is that it provides insight into underlying molecular interactions and mechanisms, which are often inaccessible experimentally, within the limits of the approximations in the models and the theory concerned. Computer simulations can be performed under conditions where it is difficult or impossible to conduct experiments, for instance, at very high pressures and temperatures. But, beyond the provision of qualitative insight, as our understanding increases one would hope to use these methods to quantitatively predict the outcome of experiments prior to, and indeed even instead of, performing them [1–3]. In this way, computational techniques should reduce time and cost in industrial processes like the discovery of drugs and advanced materials, which take more than 10 years and $2.6 billion for the former [4], and 20 years and perhaps $10 billion for the latter. Due to these potential benefits, computer-based techniques

are becoming increasingly popular among researchers from diverse backgrounds, and are adopted as routine techniques by a significant section of the scientific community. The relentless enhancement in the performance of high-end computers is another key factor accounting for the increasing adoption of computer-based methods in science over recent decades.

Given the rapidly growing popularity of computational techniques, it is all the more necessary to ensure that these techniques are reproducible [5,6]. This is essential for such techniques to be relied upon for taking actionable decisions and thereby to become a standard technique applicable in diverse applications, including industrial and clinical contexts. Here, we focus our review on the field of ligand–protein free energy calculation methods based on classical molecular dynamics (MD) simulations and biomolecular systems. A systematic account of the lack of reproducibility of many published results from *in silico* MD and an explanation for their occurrence are provided. Ensemble-based methods are the central focus of our attention since these provide the correct statistical–mechanical way in which to calculate macroscopic quantities such as free energies from microscopic dynamics. They also permit us to perform uncertainty quantification (UQ) in respect of the computed results, and underpin their verification and validation by means of statistically robust procedures. UQ is an established domain in applied mathematics and engineering but has been notably absent *inter alia* from the analysis of computer simulations performed using electronic structure and molecular simulation methods. At this time, we are witnessing unified developments in quantifying uncertainty in computer simulation across a wide range of domains including weather, climate, material, fusion, molecular and biomedical sciences [7–13].

A major goal in drug discovery and personalized medicine is to be able to calculate the free energy of binding of a lead compound or drug with a protein target. That target may be either a generic protein or, in the context of personalized medicine, a sequence-specific variant, reflecting the fact that individuals may respond differently to a given drug based on their genetic makeup. For such calculations to be useful for real-world applications, for example, in drug discovery and clinical decision making, the predictions must be arrived at rapidly, preferably within at most a few hours; manifestly, they should also be accurate and reproducible.

The free energy of binding, also known as the binding affinity, is the single most important initial indicator of drug potency, and the most challenging to predict. It can be determined experimentally by a number of methods, of which measurement of half-maximal inhibitory concentration (IC_{50}) provides a semi-quantitative estimate (technically speaking, it is just a 'proxy' for the true thermodynamic binding affinity), while biophysical techniques such as isothermal titration calorimetry (ITC) and surface plasmon resonance (SPR) are quantitative, albeit much more time consuming. It should be noted that even the more quantitative measurement methods like ITC are well known to have problems yielding accurate and precise thermodynamic parameters [14]. Indeed, one of the more surprising things about experimental binding affinity data, given their apparent importance in drug discovery, is the extent to which they are reported in widely used databases without any mention of either the measurement method used, or the associated measurement errors.

Alternatively, one may seek to calculate the binding energy theoretically. Here, methods drawn from computational chemistry offer a route forward; these are primarily based on *in silico* MD, for which several approaches to determining the free energy are possible. Methods that rapidly predict binding affinities are preferable in the context of personalized medicine and drug discovery but, as we shall see, there is a trade-off between computational cost, accuracy and precision. As a consequence of the conflation of experimental methods and their unknown error distributions referred to in the preceding paragraph, these computational approaches are hindered in a number of ways. Nonetheless, by advancing the accuracy and precision of theoretical methods one may expect to encourage more care to be taken in reporting similar attributes of experimental binding energies.

The widespread use of molecular simulation for free energy calculations, especially in the field of pharmaceutical drug development in the last few years, is now placing a premium on our ability to deliver *actionable* predictions to academic, industrial and clinical communities. For knowledge to be actionable in the current context means that the predictions are accurate, precise and reproducible, and are made on time scales that are sufficiently rapid to be used in a decision-making context. The most familiar example of actionable predictions arises in weather forecasting, as well as climate science, where ensemble-based methods play a central role [15,16]. People wish to know tomorrow's weather today, not tomorrow let alone in three months' time. Having a reliable probabilistic prediction prior to an event taking place is extremely valuable and arguably represents the apotheosis of the scientific method in action. There is a growing awareness of the importance of making actionable predictions for a range of real-world problems, reflected in recent literature covering a range of disciplines including natural disasters, climate change and medicine [17–20]. This review aims in part to enhance awareness of the issue in computational chemistry and molecular simulation in particular.

2. Dynamical systems, ergodic theory and statistical mechanics

All the free energy methods we shall describe are based on the use of classical MD. The dynamical observables, G, are calculated as macroscopic averages, which are given by ensemble averages, denoted $\langle G\rangle_t$. Thus

$$\langle G\rangle_t = \int G(x)\rho_t(x)\mathrm{d}\mu,$$

where x denotes the $6N$ phase space variables, μ is the invariant measure associated with it and N is the number of particles in the system. The ergodic theorem is commonly invoked within the domain; it states that, in the long time limit, a single trajectory generates a time average of a dynamical observable, $\langle G\rangle_t$, that is identical to its ensemble average $\langle G\rangle_{\mathrm{eq}}$:

$$\langle G\rangle_{\mathrm{eq}} = \lim_{t\to\infty}\langle G\rangle_t = \lim_{t\to\infty}\int G(x)\rho_t(x)\mathrm{d}\mu = \int G(x)\rho_e(x)\mathrm{d}\mu,$$

where ρ_t and ρ_e are respectively the $(6N+1)$ dimensional time-dependent and equilibrium probability distribution functions defined on the phase space. ρ_t satisfies the Liouville equation [21]. A time-independent state is asymptotically

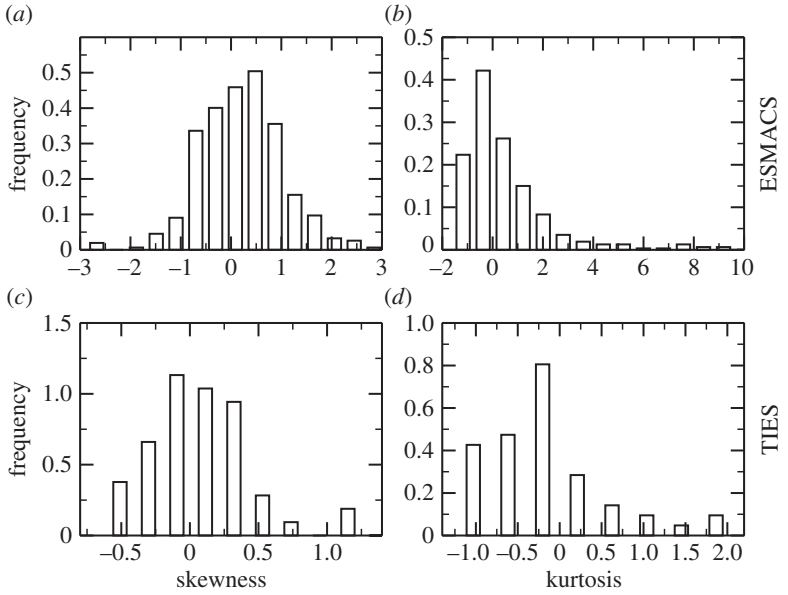

Figure 1. Non-Gaussian properties of equilibrium distributions from ensemble MD simulations. The Fisher–Pearson coefficient of skewness (*a*,*c*) and the Fisher kurtosis (*b*,*d*) for distributions of predicted binding free energies using ESMACS approach (*a*,*b*) and free energy differences from TIES calculations (*c*,*d*). The ESMACS studies consist of approximately 400 ligand–protein complexes, and the TIES studies include alchemical mutations of 50 pairs of ligands. The distribution of calculated binding free energies for each ligand comprises 25 independent replicas within an ESMACS simulation; the distribution of binding free energy differences for each pair of ligands comprises 20 or 40 replicas in a TIES simulation. The distributions are skewed both left (negative values of the skewness) and right (positive values). For the distributions with kurtoses less than 0, the distributions have short and thinner tails, and are more flat-topped than the normal distribution would predict. For the cases with kurtoses greater than 0, the distributions have a heavy tail, usually at the right for ESMACS (*b*) because the binding free energy has a lower bound (on the nanomolar or picomolar level). The large kurtosis values in these studies indicate that such simulations produce significantly more 'outliers' than one would anticipate were the statistics to conform to a normal distribution.

approached if the dynamical system possesses an equilibrium state such that

$$\lim_{t\to\infty} \rho_t = \rho_e.$$

To be ergodic, a system must pass through every possible point in phase space on the energy shell (in the microcanonical ensemble). The probabilistic description of the dynamical behaviour should be invoked; although usually stated as being equivalent to the deterministic Newtonian, trajectory based, formulation of classical mechanics its conceptual basis is quite distinct and admits the inclusion of statistical mechanical concepts which are lacking in the trajectory-based approach [21]. The problems surrounding the reproducibility of the method are rooted in the instability of the MD trajectories that underpin it, rendering them increasingly inaccurate as time evolves [21]. Perhaps surprisingly, this fundamental issue is something that has been frequently overlooked in the literature. It is perhaps not well known [21–27] that many complex systems, including all those to which statistical mechanics is applied in order to calculate equilibrium states and their properties, exhibit extreme sensitivity to initial conditions. Briefly, in the ergodic hierarchy of dynamical systems, systems which approach and reach equilibrium must be at least mixing [21]. Neighbouring trajectories—the solutions of the Newtonian equations of motion—in such systems, no matter how close they are initially in phase space, diverge exponentially fast with the passage of time [21]. Under such circumstances, the notion that we can, even in principle, specify by some experimental procedure the initial conditions for the time integration of the dynamics is undermined. Instead, we are obliged to formulate the approach to equilibrium in probabilistic terms [21,28,29], which is to say in mathematical language based on measure (also known as Lebesgue) theory. That theory holds almost everywhere, except possibly for a set of zero measure. But everything we can compute is based on a very small subset of the computable numbers, the IEEE floating-point numbers, both of which are sets of zero measure, the latter being a very small subset of the rational numbers (they are dyadic numbers, that is numbers with power of two denominators). Representing behaviour of dynamical systems using floating-point numbers may lead to the omission of a considerable amount of the structure of a dynamical system as has recently been pointed out [30]. Therefore, there is the likelihood that the calculations performed on modern digital computers, using the IEEE floating-point numbers, may not always correctly describe the probabilistic properties of chaotic dynamical systems [30]. The ramifications of these limitations remain to be fully understood but our expectation is that they may typically contribute systematic errors of order unity to the expectation values computed on digital computers. Given that, for the computation of free energies in this way, we are calculating expectation values which are estimated by 'sampling' from what one hopes to be the true probability distribution, this sampling must be performed in a manner that is sensitive to all possible sources of uncertainty.

Extensive studies we have performed in recent years confirm that MD models indeed exhibit sensitivity to initial conditions [24–26,31]. From our investigations, we find that the frequency distribution of observables as it emerges from the members of an ensemble exhibits deviations from the standard Gaussian profile anticipated on the basis that the variables are independent of one another, as one assumes in conventional statistics [32]. 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 (figure 1*a*,*c*). The majority of the distributions have positive kurtosis, meaning they are heavy-tailed relative to a normal distribution (figure 1*b*, *d*). This is at first sight unexpected—it has not previously been

reported in the context of classical MD—until it is recognized that these systems all display chaotic behaviour as well as long-range interactions; the underlying nonlinearities in the dynamics are what accounts for both the presence of chaos and non-Gaussian statistics. In addition, the distribution of experimental binding free energies might not be Gaussian. The phenomenon is well known in turbulence: there it is caused by very long-range hydrodynamic interactions mediated by energy dissipation. The reason for the presence of non-normal statistics in biomolecular systems at equilibrium comes from the fact that here too we are dealing with the infinite range interactions mediated by Coulomb forces. 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, leading to the more frequent occurrence of outliers than otherwise would be expected. Graphs of theoretical versus experimental data will not produce ideal linear plots in which all points converge closely on the 45° line. Instead, those working in this domain should expect to observe many more so-called outliers as a consequence of the natural behaviour of these systems. Although no explicit mention is made of the fact by the authors, Knapp *et al*. [33] is replete with figures displaying skewed (i.e. non-normal) distributions of geometrical quantities emanating from large ensembles of protein MD trajectory data. Statistical tools like bootstrapping and linear regression do not assume any underlying distribution. Therefore, in principle, they should be applicable to non-normal observables, such as MD-based free energies. However, their quantitative reliability is debatable for non-normal distributions in the absence of sufficient samples [34,35]. This further adds to the necessity of performing ensemble simulations for MD-based methods. Additionally, estimators like median-of-means allow heavy tails, are robust to outliers in the data and hence may be used to estimate means and variances for such distributions. As soon as this behaviour is apprehended, one understands why the outcome of single simulations is in general not reproducible, and may lead to false-positive conclusions [33]. It has an anecdotal quality. The next person who studies the same system is likely to obtain very different results (as indeed one observes in the literature when, for example, one paper reports the observation of a conformational change while another does not, from the very same system [21]).

Our findings serve to underscore that we need a statistical theory of MD simulation, in the same way that there is an established statistical theory of turbulence [36]. Individual trajectories are not robust indicators of molecular behaviour, and owing to chaos long duration single trajectories lack accuracy, but we expect statistical averages over such trajectories to be knowable to high precision. In this way, we can distinguish random from systematic errors, the former arising due to the chaotic nature of the dynamics, the latter to errors in things like the force field parametrizations employed. Without first correctly handling the stochastic errors, it is not possible to assess correctly the nature/size of the systematic errors. Our recent work [30] shows, however, that even statistical averaging can produce hitherto unexpected systematic errors, caused by the fact that the IEEE floating-point numbers are a poor representation of the real numbers. We have shown this recently for the case of simple dynamical systems for which the equilibrium probability distribution is known exactly [30]. For the systems studied by practitioners of MD and turbulence, such probability distributions are not known and must be assessed by sampling methods. Based on our studies of very simple ergodic systems, the errors accruing from the use of floating-point numbers are, in specific cases, likely to be catastrophically large; in others they are more insidious, in that simulation results may seem correct but will contain errors of order unity. In the most favourable cases they will contain errors that are small, but nonetheless much larger than machine precision.

Putting those fundamental limitations of floating-point numbers aside, the calculation of observable stochastic quantities—which are essentially expectation values—proceeds through an ensemble approach, in which a set of independent MD simulations, referred to as 'replicas' in statistical mechanics, is performed and averaged over both time and the members of the ensemble (see details in the following subsection). The key feature of ensemble simulation is the use of ensemble and time averaging [21]. 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 difference to the expectation values calculated. This is very different from calculations using a few repeats of a single MD simulation [37,38] including replica exchange [39], which do not permit reliable estimation of errors. Here, we describe our procedures for performing rapid and reproducible binding affinity calculations, and present the results for a number of ligand–protein cases. The approach is scalable: the throughput of results depends primarily on the size and speed of the available computer. Recent work [40] on both the calculation methods and the cyberinfrastructure environment could transform drug design by supporting accurate and rapid calculations of how strongly compounds bind to target molecules, a high-performance application that can scale to the largest supercomputers available in the world today and well beyond [41].

2.1. Making molecular dynamics simulations reproducible

Scientific results are by definition supposed to be reproducible, that is they should be independent of who conducts the study. Within the scientific community, there is a somewhat confused terminology addressing the confirmation of the correct measurement or phenomena, which includes such terms as reproducibility, replicability and repeatability [6,42]. We use the term 'reproducibility' in the context of this review to refer to the ability of a method, be it experimental or theoretical, to yield the same results when repeated *by oneself or others*, *with or without variation in its implementation* including the software and the hardware employed. Such reproducibility, which necessarily has a statistical nature, is essential if a technique is to be used, for example, in a medical context to treat human beings, where there are stringent regulatory requirements. But it is also fundamental to the development of reliable scientific methodologies in much wider contexts [43].

A method cannot be reliable if it does not yield the same result when performed by oneself, let alone others. Indeed, the lack of reproducible results in the published literature is a widespread concern in the scientific community [6]. Measurements are not perfect. They always contain errors. So too for a theoretical result. The issue was recently highlighted by a survey conducted by *Nature* which found that more than 70% of researchers failed to reproduce another researcher's results, while more than half were unable to reproduce their own [44]. In the case of experiments, non-reproducible results can be an artefact of factors ranging from the incorrect use of

chemicals, an insufficient number of samples, fluctuations in the environment and variations in the experimental setup, to data dredging and *a posteriori* hypothesis generation, not to mention conclusions influenced by conformational bias, conflicts of interest, selective reporting or, worst of all, misconduct. In the case of computer-based methods, the reasons may also reside in the theory or the model used, the extent of convergence of the calculations, the reliability of the software, the adequacy of the floating-point representation of the real numbers [30] and so on [27].

In this review, we focus on the convergence, reproducibility and reliability of observable properties obtained from MD simulations. Although it was recognized more than two decades ago that one-off classical MD simulations do not generate consistent protein conformations [45,46], systematic investigation as to how to make these calculations reproducible has not been performed until recently. This is a reflection of the history of how ensemble methods became adopted in weather and climate forecasting: pioneered some 25 years ago, their initial introduction encountered considerable resistance from established workers in the domain but they are now the standard method by means of which probabilistic weather forecasts are made on a daily basis [16]. In the molecular simulation domain, one notices similar reluctance to embrace the probabilistic approach. Until now, one-off simulations remain a very common way of performing MD studies (see 'Application of free energy calculations' section below).

More specifically in the field of MD-based free energy calculations, the variation in the calculated free energies based on independent simulations was investigated systematically by Sadiq *et al.* [24] and by Genheden & Ryde [47] using MMPBSA and MMGBSA methods, respectively. The estimated free energies from two independent MMPBSA calculations of the same molecular system can vary by more than 10 kcal mol^{-1} in smaller molecule–protein complexes [26,31,47], and by up to 43 kcal mol^{-1} in larger and/or more flexible ligands bound to a protein such as the peptide–MHC (major histocompatibility complex) systems [25]. The underlying reason for such variations between independent MD simulations is due to the extreme sensitivity of a dynamical system to its initial conditions [21,48].

Although these studies [21,24,45–47] employed various methods to calculate diverse observable quantities, the common conclusion was drawn that multiple short MD simulations provide substantially better sampling than a single long MD simulation. This does not invalidate the ergodic theorem, it merely indicates that the time scales over which MD simulations are run are nowhere near long enough to fulfil its requirements. It should be noted that these studies investigated systems which were at thermodynamic equilibrium; for behaviour that requires long time scales to happen (such as conformational changes), including also the important case of non-equilibrium systems, we must use ensembles for both long and short simulations. Further investigation has been undertaken in the last few years, which has led to approaches such as enhanced sampling of MD with an approximation of continuum solvent (ESMACS) [25], velocity-induced independent trajectories (VIIT), solvation-induced independent trajectories (SIIT), conformation, rotation and protonation-induced independent trajectories (CRPIIT) [49] and methods-induced independent trajectories (MIIT) [50]. Similarly, in the case of alchemical methods, Lawrenz *et al.* [38] introduced a method called independent trajectory thermodynamic integration (IT-TI) which employs multiple, independent TI calculations and yields more accurate free energy changes. More recently Bhati *et al.* [51] published a method called thermodynamic integration with enhanced sampling (TIES) which employs an ensemble of independent MD simulations in combination with the concept of stochastic integration to yield accurate and precise free energy predictions. While most of the aforementioned approaches only make use of multiple separate simulations and/or independent trajectories but do not systematically assess the statistical or the statistical mechanical significance, ESMACS and TIES exploit the statistical mechanical concept of ensembles and the connection to ergodic theory to quantify uncertainty and obtain reproducible results from MD simulations in a systematic and theoretically well-grounded manner [21].

2.2. Sources of error in classical molecular dynamics

As noted above, there are two sources of error accruing in MD simulations, due to systematic and random sources. The systematic errors originate in things like the imperfect methods, models and calibration of simulations. Biases of protein force fields towards different secondary structure types [52], for example, will consistently populate either helical or sheet-like structures from independent simulations. When the cause of such systematic errors can be identified, it can be reduced or even eliminated, as shown in recent simulations with state-of-the-art force fields [53].

Random variation, also called system noise or stochastic error, on the other hand, has a different origin. It is caused by the chaotic nature of classical MD. Given the 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 of the simulations [21]. The difference produced by two simulations introduces a level of variation which can be larger than the quantity of interest, making the results practically useless.

In order to get a full grip on uncertainty in MD simulations, one needs to be able to identify the systematic and random components contributing to the errors. We would like to assess the variation in results arising from the inaccuracy inherent in the molecular models including the choice of force field. This has not been convincingly addressed to date, since for the most part the UQ due to the intrinsic random error in the simulation trajectories is only now being nailed down. The overall quality of a computer-based simulation study can be assessed by a relevant set of verification, validation and uncertainty quantification (VVUQ) methodologies and associated tools [54,55]. Validation is about comparing the results of our models with experiment (or high-quality reference benchmark/theoretical results) while verification is concerned with ensuring that the quantities we are calculating from the algorithms and software are themselves being computed correctly. UQ reports the error in the calculations. It is timely to address these issues now and one purpose of this review is to encourage such investigations (see for example https://www.vecma.eu/). One way to approach these issues is through sensitivity analysis in the first instance—which term or terms in a force field or other choices made, such as cut-off distances, size of simulation cell, etc. lead to the greatest sensitivity, for example in the calculation of the free energy of binding of a ligand to a protein?

2.3. Ensemble averaging

In order to address the aforementioned problems, it is necessary to perform ensemble simulation, requiring a break-away from the traditional practice of performing one-off MD simulations. The practice of one-off simulations, based on an *ad hoc* appeal to the ergodic theorem discussed earlier, has been at the basis of most publications in the field since the dawn of MD which was originally introduced by Alder & Wainwright in the late 1950s [56]. We focus here primarily on free energy methods which have been used for decades to study the binding affinities of ligands to their target proteins [57].

The results from ensemble simulations are designed by construction to be accurate, precise and reproducible [21,24–26,47,50,51,58–61]. It should be noted that the term 'accuracy' refers to the closeness of the results to the corresponding experimental values and it is largely dependent on the limitations of the force field employed. In addition, given the vast number of nodes, cores and accelerators on modern high-performance computers and available automated workflows (see the 'Distributed computing approaches to enhance sampling' section below), all of the replicas can be run in parallel and hence an ensemble simulation can be run in the same wallclock time as needed for computing a single replica [25,51]. This leads to rapid predictions informed by high-quality error estimates, which is essential for free energy methods to have an impact beyond an academic setting. The appropriate number of replicas and the duration of the simulation are parameters dependent on the system under study and the calculational method selected; they depend on the extent of stochasticity, that is the fluctuations, within individual stages in a calculation comprised of multiple steps, and the level of precision desired, although general guidelines are available [21,25,26,51,61]. Thus, for example, within a relative binding free energy calculation using a thermodynamic cycle, the alchemical leg for the ligand–protein complex requires more replicas than that for the ligands to realize the same level of precision [61]. The errors in each individual step decrease as the inverse of the square root of the number of replicas, $1/\sqrt{N_{\text{replicas}}}$. Varying numbers of replicas are required to achieve a desired level of precision; for example, of order 25 replicas are typically required for ESMACS studies [25,26], 5 replicas for each λ window within a TIES relative binding free energy calculation [51], a combination of 5 and 10 replicas during the multiple steps required to compute absolute alchemical binding free energies [61] and as many as 40 replicas for some graphene-related MD studies [62]. Many ensemble simulations only vary the velocities in the initial conditions of the replicas [24,25,51,61]; in some cases variations in the initial spatial coordinates are also required [63].

Considerable effort has been invested in the development of so-called 'enhanced sampling protocols' in order to improve phase space sampling [64–66], including metadynamics, a method for accelerating rare events in simulated systems [67]. Among these, the most popular in the case of biomolecular simulation is the Hamiltonian-replica exchange (H-REMD) [68] and its variants—replica exchange with solute tempering (REST2) [69] and FEP/REST [70]—which run multiple concurrent (parallel) simulations and occasionally swap information between them to improve sampling. A molecular system subjected to these parallel simulations has a common configuration space; the simulations sample the microscopic states with different probabilities because of the differences in their Hamiltonians. For a given set of simulation samples, different free energy estimators can be applied with varying reported accuracies and precisions [71]. One of the free energy estimators is called the multistate Bennett acceptance ratio (MBAR) [72] which has become increasingly popular of late. MBAR makes use of all microscopic states from all of the replica exchange simulations, by reweighting them to the target Hamiltonian.

Free energy calculations had rarely been used seriously in drug development projects until recently when Schrödinger Inc. released their 'FEP+' simulation software for relative free energy calculations [73]. With the improved technology and the availability of graphical processing units (GPUs), FEP+ has made a significant impact in the pharmaceutical industry within its domain of applicability [74]. FEP+ is employed in a shrink-wrapped and very easy to use manner, being wholly proprietary and directed at commercial users. Unfortunately, it promotes the uncritical use of one-off simulations. Merck recently published a large-scale study using FEP+ and discussed several challenges in its application within the drug discovery process [75]. However, the authors did not perform ensemble simulations to confirm the robustness and reproducibility of their findings which therefore have only a provisional status. In an attempt to gain a handle on errors in these calculations, FEP+ recommends the use of closed thermodynamic cycles of transformations performed by one-off simulations in order to detect hysteresis and, indirectly, to assess convergence. While this method can indicate that convergence has not been reached, a small error reported in cycle closure convergence does not guarantee convergence. This is because, while a hysteresis value of 0 from such a closed cycle is a necessary condition, it is by no means sufficient. Hysteresis certainly merits close attention, as indeed we do within TIES, for example, when comparing results from simulations running several 'forwards' and 'backwards' transitions [76]. In such cases, this amounts to a form of ensemble simulation in which replicas with different initial and final states are used.

It is often claimed that the implementation of an enhanced sampling protocol such as REST2 [69] and the use of the free energy estimator MBAR [72] can overcome the problem of non-reproducible results. This is not the case. Application of REST2 may make a ligand drift away from its stable binding position, and lead to deteriorated free energy predictions [77]. In recent work, we performed free energy calculations using FEP+ [73] which *de facto* implements REST2 and MBAR. Up to 3.9 kcal mol^{-1} variations were observed from 30 independent simulations, much larger than the MBAR errors reported for individual FEP+ calculations [78]. Other studies have also found that bootstrap analyses from repeated simulations provided a more realistic uncertainty estimate than MBAR [79]. It is clear that such 'enhanced sampling' methods are not an alternative to the use of ensemble simulations [61]: they too must be ensemble averaged [61,77,78].

3. Common methods for free energy calculations

Molecular recognition [80] is central for many physical, chemical and biological processes. Accurate prediction of the binding affinity of a guest molecule with its host is an

important goal in host–guest chemistry and has an essential role in biomolecular signalling and pathways. A guest is often a compound with low molecular weight, while a host is usually a larger molecule which encompasses the guest. Guests are the so-called ligands we consider here, which are molecules that bind reversibly and specifically to a biomacromolecule (a protein in the context of this paper) and alter the latter's activity. MD simulation provides a tool to collect microstates of the biomolecule of interest, and has been applied to get the macroscopic thermodynamic properties, such as the free energies, from the ensemble of microstates.

3.1. Free energy of binding

The binding affinity is the change in the free energy associated with a binding process. The magnitude of the binding affinity is a measure of how strong the interaction is between the ligand and the protein, and hence it is often directly related to the potency of the ligand. Therefore, its measurement is of importance in the fields of drug design and personalized medicine. It can be used as a virtual screening tool in drug design or as a clinical tool to tailor a patient's medication based on his/her genetic makeup. Computer-aided drug design (CADD) is an extremely active field of research [81]. In addition, rapid and accurate binding affinity predictions can be useful in health-related applications like the design of medicines with reduced side-effects and drug resistance [82]. Thus, the use of *in silico* techniques to predict binding affinities has grown immensely in the last few decades [83].

Reliable binding affinity predictions need to be made on time scales shorter than experimental ones in order to have a real impact in drug design and personalized medicine [82]. Therefore, the time to solution is another important factor influencing the applicability of computational methods in real-world scenarios. This is especially crucial when considering the development of a typical prescription drug which, as noted, often takes around a decade (and costing $2–3 billion) to get to market [4]. The ensemble approaches described above fulfil these requirements given the availability of sufficient computing resources (for more details see the 'Application of free energy calculations' section).

3.2. Methods for free energy calculation

The methods with most potential, in order of increasing level of molecular resolution, are listed below. A higher level of resolution ought, in principle, to lead to higher accuracy, although this is not necessarily true because of the quality of the theory employed and the way in which the calculations are implemented [27].

(i) 'Informatics' based approaches which are usually the output of docking studies in combination with so-called 'machine learning' [84–87];
(ii) linear interaction energy (LIE) methods [88];
(iii) molecular mechanics Poisson–Boltzmann surface area (MMPBSA) and molecular mechanics generalized Born surface area (MMGBSA) methods [89] based on invoking a continuum approximation for the aqueous solvent to approximate, e.g. electrostatic interactions following all-atom MD simulations; and
(iv) alchemical methods including thermodynamic integration (TI) and free energy perturbation (FEP).

Machine learning (ML) is currently gaining a lot of traction within the pharmaceutical industry. The current approach is to seek to invoke ML to generate candidate compounds and to rank congeneric compounds [90]. There are now many start-ups and small companies that offer such 'AI' (artificial intelligence) based approaches to drug discovery; this is being done to generate lots of candidate compounds, both virtual and real. The predictive performance of ML methods for binding affinities, however, is sensitive to the quality of the ligand–protein structures which are usually generated using docking methods. There are claims that ML can achieve 'chemical accuracy', meaning ±1 kcal mol^{-1} in energy predictions. Indeed, it has been shown that, when the most relevant high-quality data are used for training, ML algorithms can generate accurate binding affinity predictions [91,92]. This is still hotly contested in real-world situations due to a number of shortcomings of such approaches [1,2,93]. These arise from some obvious built-in assumptions of all ML algorithms. The key one is the assumption that relationships between points in ML data space are smooth (continuously differentiable), so they interpolate smoothly between gaps in the state space. This may or may not be valid, depending on each and every case under study; when invalid, however, its predictions will fail badly. Thus, for example, when the free energy changes more or less discontinuously with molecular structure, as it does for the case of free energy cliffs [94], there is no way such an ML algorithm will in general be able to spot such phenomena. It could only do so if the coverage of the state space were exceptionally dense, implying that the data upon which it has been trained would need to be enormous. A related generic problem is the well known 'curse of dimensionality': for a state space of dimensionality N, the required quantity of training data grows as an exponential function of N, so that there is no chance of acquiring sufficient data to get close to densely populating the state space of a complex system with representative examples. ML is, at root, nothing more than glorified curve fitting; and equipped with so many thousands of adjustable fitting parameters, it is no wonder that it may appear to fit the data it has been trained upon well. In general, however, this leads to overfitting, meaning that it then often fails spectacularly but unexpectedly when asked to make predictions for previously unseen data [1,2]. Lacking any significant explanatory power, it is hard to figure out what has caused the poor performance.

More accurate experimental and computational chemistry studies are needed to provide correct binding poses and binding affinities in conjunction with it. Combinations of ML and MD are currently being used, for example, to search for appropriate evolution of MD simulations through various 'on-the-fly learning' processes [95,96]. Because of the high compute intensity of MD calculations, ML is being increasingly used as a 'surrogate' in order to replace that expense with something less costly and time consuming. It is hoped that combinations of ML and MD [97] will enable the virtual screening of colossal numbers of compounds, and to focus only on a small subset of those virtual compounds with more computationally expensive free energy calculations. This, in turn, should ultimately lead to much more limited effort and cost expended on the actual synthesis and testing of candidate compounds for subsequent drug development.

The LIE approaches usually generate worse relative binding affinity rankings and considerably larger uncertainties than MMPBSA and MMGBSA [98]. In addition, the scaling factors for the electrostatic and the van der Waals interactions

in LIE approaches are still a matter of discussion, and the quality of predictions from the approaches is frequently reported to be system-dependent [98].

Ensemble-based approaches centred on (iii) MMPBSA/MMGBSA and (iv) alchemical methods are the focus of our attention in this review. Although MMPBSA/MMGBSA means many different things in the literature, when we refer to it here we mean the full determination of the free energy of binding from either a one-, two- or three-trajectory method: it includes both the configurational entropy and the association free energy [26,99], and—where appropriate—the adaptation energy [25,59,60]. It may be invoked, in principle, to any inhibitor–protein complex, although caution should be applied when truly diverse datasets are handled [100,101]. ESMACS is the name we give to this protocol when it is run in an ensemble-based form, with all these options available to select from. Recent ESMACS publications [25,59,60,100,102], for example, investigate drug-like small molecules bound to therapeutic targets, including G protein-coupled receptors (GPCR), the most frequently exploited drug target class, as well as biological substrates (9-mer peptides) bound to the major histocompatibility complex (the p-MHC system) of central importance in immunology. The peptides in the latter are much larger than common small-molecule drugs, and have widely varying structures and electrostatic charges. The methods have also been used to study protein–protein interactions [103,104]. ESMACS is thus well suited for use in the initial hit-to-lead activities within drug discovery.

The alchemical methods have a more restricted domain of validity: they are applicable mainly to estimating small relative free energy changes for structures (drugs or proteins) which involve relatively minor (perturbative) variations. As such, the methods are most relevant to lead optimization following the identification of promising lead compounds. When changes in the net charge arise, TI and FEP methods encounter specific difficulties owing to major adjustments in long-range electrostatic interactions [105]; charge correction approaches are required to take into account the artefacts of the electrostatic potential energy introduced by the finite size effects [106]. A recent paper by Wang *et al.* [73] employing FEP has attracted significant attention, as it purports to provide a reliable route to the prediction of binding free energies. However, it has the same restricted scope as it seeks to compute free energy differences between similar ligands bound to a protein. The approach advocated has been to run single simulations, without paying any attention to the stochastic nature of the quantities calculated. A recent study furnishes an estimate of the reproducibility of TIES and provides a reliable method for UQ for both relative and absolute binding free energy (ABFE) calculations using alchemical methods [61,78]. While equilibrium simulations are commonly implemented in alchemical methods, there are also non-equilibrium approaches which can generate comparable results with these obtained from equilibrium simulations [107–111]. The accuracy of such approaches is quantified by comparing the predictions with experimental data which, as previously discussed, all too often have no associated errors. Compounding this, there is frequently no reporting of the variations arising from the predictions provided by the calculational method, leaving the uncertainty associated with the protocol largely unquantified.

While ESMACS (and LIE) is an ABFE method and TIES/FEP+ are relative free energy methods, there exists an alchemical ABFE method which can be used to estimate binding affinities, which we now describe. It is the equivalent of TIES/FEP+ when one of the drugs involved is replaced by nothing, in both bound and unbound states. The calculation method is called double annihilation, first proposed three decades ago [112]. A series of nonphysical steps are involved in the calculation; the free energy changes for each step are calculated by a combination of alchemical and analytical methods [61,79,112]. The processes of decoupling/coupling the ligands from/to the environment involve large changes in phase space, of which the calculated free energy changes exhibit large fluctuations. Owing to its extreme compute intensity and intrinsically large uncertainties, the method has until recently not been applied to pharmacologically relevant proteins in any significant manner. Ensemble approaches render ABFE much more reliable, and reveal that in this compute-intensive multistep calculation, the various steps require different ensemble sizes to attain the same high level of precision [61]. The ABFE calculation is by far the most expensive of the methods we discuss here. Compared with the pair-wise comparison of the relative binding free energy calculations, the advantage of ABFE approaches is that their results can serve as a reusable library to which calculated ABFE results for other ligands can be compared and added.

The use of these approaches is not mutually exclusive but indeed can be even more powerful when performed in tandem. For example, a combination of endpoint and alchemical methods has been used to accurately predict protein–ligand interactions for a membrane transporter [113]. In the currently ongoing COVID-19 pandemic, the computer-aided drug discovery market has experienced a boost, in which the aforementioned approaches have been extensively applied, separately or jointly, to find novel drug candidates and to reposition existing drugs. We ourselves are currently participating in a large scale collaboration in which ML, docking, endpoint and alchemical approaches are applied interactively to find promising drug candidates from data consisting of billions of compounds. The most attractive drug candidates are subsequently being studied experimentally, with some under consideration for inclusion in possible clinical trials.

4. Ensemble-based simulation approaches

Over the past 20 years, all these methods have been subjected to substantial criticism for a wide range of reasons, mainly due to their lack of accuracy and reproducibility, and in the case of (iv) their long turnaround time. Two distinct but related problems contribute to the issue: conformational exploration and precise sampling. The usage of ensemble approaches is increasingly widespread within a broad range of MD studies, for sampling 'rare events' including protein folding and ligand binding kinetics using ensemble dynamics [114], weighted ensemble [115] and splitting methods [116], for predictions of residence times using steered MD [117] and random acceleration molecular dynamics (RAMD) simulations [118], and extending now from all-atom to multiscale and coarse-grained studies [62,119,120].

4.1. Ensemble-based conformational exploration

Many biological processes occur on time scales which are difficult, if actually possible, to access by atomistic MD

simulations. Such processes usually go through a complicated free energy profile which can be simplified into local minima and transition states. The former normally trap a system for very long times, while the latter can only be accessed rarely and transitorily [9]. Transition path sampling [55] is one common approach to investigate the transition paths connecting different minima, in which accelerated MD approaches such as metadynamics [67], steered MD [117] and high-temperature simulations are first used to construct an overall free energy landscape, followed by ensemble simulations from putative transition states to generate a transition path containing information on the mechanism and kinetics of the process [121]. The ensemble simulations consist of many relatively short runs sampling regions between different minima but do not need to spend a long time in any specific minimum. A converged free energy landscape can then be reconstructed once all of the regions have been fully explored [121], if the weighting factors can be correctly assigned to each conformation [61]. A widely used approach is to construct a Markov state model (MSM) [122,123] for the description of biological processes such as ligand binding and protein folding. Large scale ensemble simulation needs to be run to adequately sample the entire configurational space, which consists of a vast number of individual MD simulations.

Among ensemble simulation approaches used for studying long time scale events by accessing transition states are ensemble dynamics [114], weighted ensemble methods [115] and multilevel splitting along with their variants [116]. These methods differ from the ones described above in that they do not involve any external force or biasing potential; nor do they involve heating the system of interest. Rather they exploit the fact that MD simulations, being chaotic, are extremely sensitive to their starting conditions and enhance sampling of otherwise inaccessible states by running large numbers of short independent MD simulations varying only in their starting conformations. The ensemble dynamics method has been successfully used to study protein folding of a large number of systems in the last couple of decades [124]. It involves replacing a single long simulation by an ensemble of shorter simulations with a cumulative simulation time of up to microseconds. The fraction of simulations capturing folding or a conformational change of interest is used to infer the probability of such events. The weighted ensemble method involves partitioning the phase space into several regions and initiating an equal number of concurrent 'walkers' in each of them [115]. The regions are maintained at equally populated levels by adjusting the number of walkers in each after regular intervals. This permits sampling of rare events that would otherwise get underpopulated with walkers and hence sampled less. The basic idea behind the multilevel splitting method is to discard trajectories that drift away from the region of interest in the conformational space while focusing on those that get closer to it [116]. To this end, a 'reaction coordinate' is defined that is used to monitor the progress of a trajectory and to measure its closeness to the desired rare event. This allows one to dedicate the majority of computational effort to sampling the region of interest instead of the initially much vaster phase space. It should be pointed out, however, that although all these methods use ensemble simulations, they do not do so to ensure reproducibility, but only as a means to capture and observe interesting transitions. Thus, the application of UQ in these contexts still remains wide open.

4.1.1. Ensemble-based docking

Ensemble methods have been used in docking studies to accommodate the flexibility of target proteins, in which an ensemble of structures can be generated from explicitly solvated MD simulations, against which the screening of ligands is performed. It has been demonstrated that the approach generates a set of top hits, some of them ranked very poorly if only crystal structure data are used [125]. Our own work [126] also showed that ensemble-based docking is required to explain the preference of gatekeeper mutant EGFR (epidermal growth factor receptor) binding with gefitinib, a targeted anti-cancer drug, rather than ATP. To improve ligand–protein binding affinity predictions, multiple independent MD simulations have been applied within the LIE [127] approach. Ensemble docking approaches have been employed recently to identify small molecules which may disrupt host–virus interactions at an entry point for infection with the SARS-CoV-2 [128].

4.1.2. Ensemble-based peptide and protein folding

Since the pioneering study by Duan & Kollman [129] two decades ago, significant advances have been made in the field of peptide folding studies [130], thanks to the rapid development of simulation approaches, the availability of powerful supercomputers, and the improvement of the force fields. Using vanilla all-atom MD, the lengths of simulations for peptide (un)folding have reached the time scales in the range of microseconds to milliseconds [130]. The Shaw group has applied long time scale simulations, up to 1 ms, correctly folding 12 structurally diverse small proteins to their experimentally determined structures using Anton [131], a special-purpose MD supercomputer. Except (un)folding in solvent, the process of peptide binding, folding and partitioning into lipid bilayers has been successfully captured using high-temperature MD simulations [132]. Because of the time scales required for the simulation of peptide folding/unfolding, it is not surprising that many of these standard simulations use a single trajectory approach, lack accuracy and reliable error estimates, and are thus unlikely to be reproducible.

An ensemble study of a 10-residue peptide [33], designed to investigate the reproducibility of MD simulations, indeed displays the necessity for applying ensemble approaches to investigate peptide unfolding. In the study, 100 replicas were investigated with a simulation length of 3 µs each [33]. The study yet again concluded that single simulations are typically not reproducible [21]. Rather than using standard MD simulations, other approaches have been applied to sample the large conformational changes involved in peptide (un)folding, such as accelerated MD [64–66] and transition path sampling [55].

4.2. Ensemble-based sampling of restricted domains

In the study of a real biological system, it is practically not possible, and indeed not necessary, to sample extensively all regions in the configuration space. Only restricted regions of conformational space are important for the calculation of many properties of interest. The binding affinity, for example, is determined by the stable binding conformations. Accelerated MD approaches such as REST2 can help conformational sampling, but their propensity to drift from stable binding conformations degrades free energy predictions because of the lack of proper weighting factors for the conformations

explored [61]. Precise predictions can be obtained when the most relevant conformations have been extensively sampled, without the need to explore a large conformational space.

The performance of short ensemble simulations for free energy predictions depends on the quality of initial binding poses and on the time scales for the efficient sampling of local conformations. The phrase 'garbage in, garbage out' is particularly pertinent when short simulations are used: a system will not be able to escape from a poor initial configuration within a limited simulation time, even if an ensemble is employed. The initial ensemble must be close to the most relevant region of the configuration space so that the relevant phase can be sampled extensively. With carefully prepared initial molecular systems, our studies show that the protocol of running 4-ns ensemble production runs works well for all the molecular systems we have investigated [21,25,26,51,59–61,77,101]; longer simulations only provide a marginal gain in the predictions [61], and may even have negative impact if the ligands drift away from their stable binding conformations [61]. There are certainly cases where longer simulation duration is required, either to improve the poor initial structures, to sample multiple binding conformations [77], or to obtain converged occupancy probabilities of water molecules at the binding sites [101,133]. A combination of conformational exploration and precise sampling, both based on ensemble simulation, may be required in these cases.

5. Distributed computing approaches to enhance sampling

In order to make a positive impact in industrial or clinical settings, computational predictions need to be made on time scales which can compete with or preferably dramatically outstrip the duration of experimental discovery and testing programmes. Today it is possible to achieve this by making use of the methods described here. The history of biomolecular simulation has been significantly influenced by the available computational power, the development of automated workflow tools, and the evolution of distributed computing approaches in recent years. In this section, we summarize the developments we and others have made in these areas.

5.1. Hardware approaches

Karplus recalled [134] the considerable courage required to perform the first MD simulation of a macromolecule of biological interest [135], due to the very limited and expensive computational resources available in the mid-1970s. Biomolecular simulation, along with other areas of computational science, have been benefiting from the rapid evolution of computing power. In the last five decades, the performance of microprocessors has improved exponentially, roughly following Moore's law which states that the number of transistors on a chip, a rough measure of processing power, doubles about every 2 years. Even though Moore's law has come to an end in recent years, the computational power is expected to keep improving for many years with the creative and/or specialized design of chips to accelerate specific crucial algorithms. There are also special-purpose supercomputers for MD simulations: Anton [131], for example, is a remarkable computer which can achieve simulation rates of microseconds per day for biosystems with millions of atoms.

Supercomputers use essentially the same microprocessors within nodes in far greater numbers than a desktop workstation, a fundamental difference arising from the speed of the interconnects linking nodes and cores together. While a typical early supercomputer from the 1980s only contained a few central processing units (CPUs), massively parallel designs since the 1990s have driven the architecture of supercomputers and connected a much greater number of microprocessors together via fast networking interconnects. Supercomputers today consist of hundreds of thousands to millions of such cores [83]. Indeed, cores as such are seldom referred to today; the basic units are the nodes, which typically contain tens to hundreds of cores and many also include accelerators.

The advent of GPU accelerators has resulted in more and more powerful specialized processing units designed for floating-point calculations. Initially used to accelerate the creation of images in a frame buffer, such processors are now in widespread use in high-performance computing and cloud platforms in so-called general-purpose GPUs. High-performance computing systems that feature both GPU accelerators and CPU chips within individual nodes allow hybrid applications to be developed that take advantage of the processing power GPUs offer, albeit at greater financial cost. MD codes developed to run on GPUs often show very significant performance improvements compared to CPU variants.

The explosion in the growth of computing power has led computational biologists to become prominent users of high-performance computing. Concomitantly, the rise of cloud computing has made accessing such resources at scale potentially trivial for researchers in academia and industry. The pay per use models promoted by clouds mean that users can pay just for the resources that they need to achieve their research objectives, without having to engage in expensive hardware procurement and operational costs, although this is a model which often works better in commercial contexts than within academic research. Notwithstanding these comments, at the level of compute intensity required to perform many free energy calculations, cloud computing can currently become prohibitive very rapidly on cost grounds.

5.2. Software approaches

To make the best use of available HPC resources, considerable effort has been devoted to improve the scalability of software and to develop adaptive and automated workflows. The scalability of MD codes on large numbers of cores and/or nodes is an important factor in determining the size and time scales of a problem which can be studied. Most of the MD codes used today have been designed or adapted to run on parallel computer systems. NAMD [136], for example, designed for high-performance simulation of large biomolecular systems, has been used to simulate systems consisting of tens of millions atoms [137], although any MD codes with long-range interactions, which are communication bound, will not scale effectively in a strong sense to reach required time scales [138]. OpenMM [139] and ACEMD (the latter now uses the OpenMM kernels) [140], designed and optimized for GPUs, are among the fastest MD codes in terms of single GPU board performance.

The application of molecular modelling techniques to real-world problems involves a complex workflow which is extremely tedious and error prone if performed manually, especially when ensemble computing approaches are embraced. In the case of pharmaceutical drug discovery for

Table 1. Workflows to simplify free energy calculations. The workflows are designed to automate one or more of the multiple-step process of the calculation, with employment of endpoint and/or alchemical approaches in conjunction with specific force field(s) and MD engine(s). ✓: function available; ✗: function not available.

	automated steps			FE approaches				
name	build	simulation	post-analyses	endpoint	alchemical	force field	MD engine	reference
FEP+	✓	✓	✓	✗	✓	OPLS	Desmond	[73]
BAC	✓	✓	✓	✓	✓	AMBER, CHARMM	NAMD, OpenMM, GROMACS	[141]
FEW	✓	✓	✓	✓	✓	AMBER	AMBER	[142]
YANK	✓	✓	✓	✗	✓	AMBER, CHARMM	OpenMM	[143]
FESetup	✓	✗	✗	✗	✓	AMBER	Sire, AMBER, GROMACS, CHARMM, NAMD, DL_POLY	[144]
pmx	✓	✗	✓	✗	✓	AMBER, CHARMM, OPLS	GROMACS	[145]
STaGE	✓	✗	✗	✗	✓	AMBER, CHARMM, OPLS	GROMACS	[146]
Flare	✓	✓	✓	✗	✓	AMBER	OpenMM	[147]

example, anywhere from hundreds to tens of thousands of compounds may need to be screened within a few days. The number of ensemble runs is of the same order of magnitude as the number of compounds. Managing the execution of simulations and collation of output data mandates the adoption of automation techniques to make the process tractable and reduce the time to solution. The scarcity of automated software tools is one major obstacle limiting the wider application of free energy approaches in real-world problems.

In recent years, a good deal of effort has been expended to develop workflows that simplify some or all of the process of free energy calculation using alchemical approaches such as TI and FEP, and/or endpoint approaches like MMPBSA, MMGBSA and LIE (table 1). The workflows include the entire steps to plan, set up, simulate, and analyse the final results in an automated manner. As mentioned earlier, FEP+ [73] is a patented free energy calculation suite from Schrödinger Inc., designed to automate the setup and analysis of FEP. It has been promoted primarily to major pharmaceutical companies. The Amber free energy workflow (FEW) [142] is a tool to set up alchemical and endpoint free energy calculations. FESetup [144] provides the setup of alchemical free energy and endpoint approaches for a few modelling packages including Amber, CHARMM, GROMACS, NAMD, Sire and OpenMM. The small-molecule topology generator (STaGE) [146] automatically generates GROMACS topologies using force fields such as Amber, CHARMM and OPLS, and sets them up for high-throughput free energy calculations, while pmx [145] within GROMACS provides an automated framework to provide hybrid protein/ligand structures and topologies for alchemical free energy calculations. Flare [147], implemented in Cresset's structure-based drug design suite, offers a graphical user interface to automate setup, simulation and analysis of free energy calculations via interfaces to the open source packages OpenMM, Sire, LOMAP, SOMD and BioSimSpace. We have developed our own free energy workflow called the binding affinity calculator (BAC) [141,148], designed to automate the end-to-end execution of ESMACS, TIES and ABFE calculations, and to handle ensemble calculations. A relative free energy calculation usually requires a hybrid topology and coordinate files to transfer one ligand to another. An automated algorithm is desired for planning and setting up free energy calculations between possible ligand pairs, and for generating required files. In most of the workflows mentioned here, this task is handled by the lead optimization mapper (LOMAP) [149] or other similar tools. The original LOMAP code was mainly based on commercial application programming interfaces (APIs) such as ones from Schrödinger and OpenEye. A new version of LOMAP has been developed, which is based on open APIs such as RDKit; this offers the scientific community a free tool to plan binding free energy calculations.

A complete free energy workflow comprises three distinct phases: (i) the preparation phase, (ii) the production phase and (iii) the analysis phase. The preparation phase is the step which all workflows focus on, in which the simulation-ready topology and coordinate files are constructed for a biomolecular system to be studied from its raw starting structure (usually in the form of a crystal structure in PDB format). They take parameter files for proteins, ligands and other components (water, ions, cofactors, etc.) as input with the specification of a desired force field. Some workflows also generate input configuration files compatible with chosen MD engines which are used to perform the equilibration and production simulations. After the successful execution of the simulation phase, the final step is to perform the statistical analysis on the output generated by the simulations or the post-processed data.

Most of the listed workflows are executed through a command line interface (CLI), with inputs being handled through

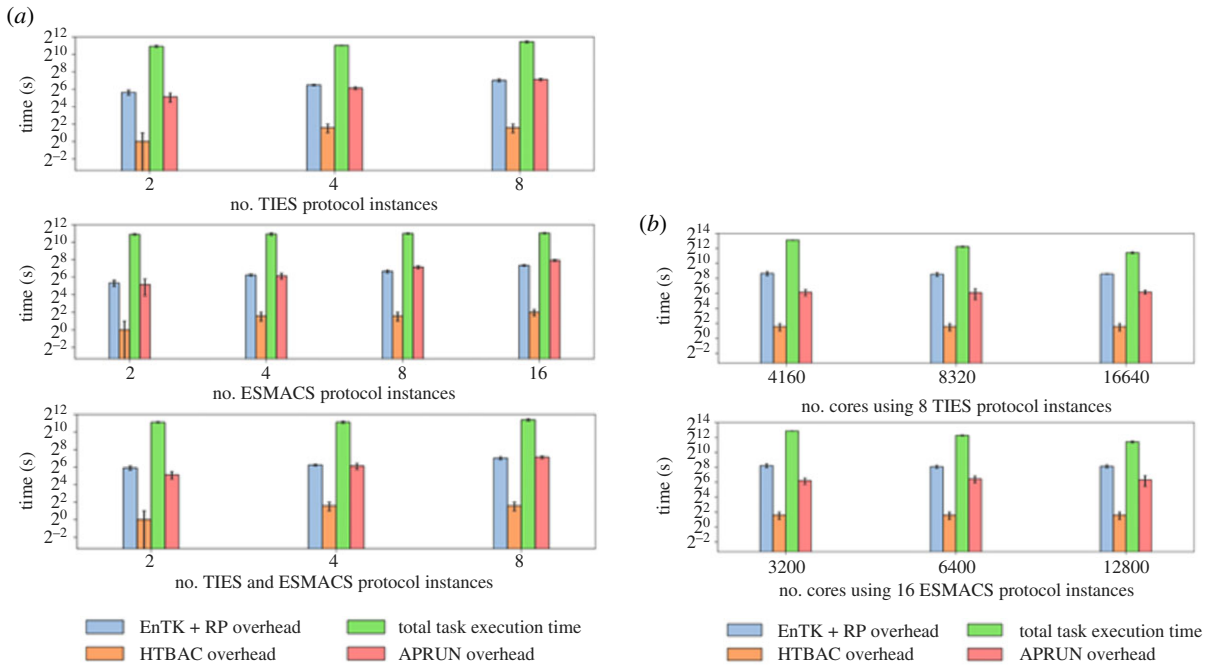

Figure 2. Weak and strong scaling of HT-BAC executed on BlueWaters, a Cray XE6/XK6 Supercomputer. (*a*) The ratio of number of protocol instances to resources is constant. Task execution time with HT-BAC, EnTK + RP, aprun overheads for TIES, ESMACS and a combination of TIES and ESMACS. (*b*) The number of protocol instances is fixed while the number of cores increases. Task execution time with HT-BAC, EnTK + RP, aprun overheads with TIES and ESMACS.

shell scripts. An API is usually available, which allows a user to access the code through their own scripts for maximum flexibility and customization. As one might expect for a commercial product targeted at end-users in pharmaceutical companies, FEP+ has a well-designed graphical user interface to ensure user-friendly operation and high-quality visualization of input and output data. It uses a range of technical approaches to improve the accuracy and throughput of calculations, based on a replica-exchange method and GPU acceleration. However, as noted above, it is proprietary so users cannot access the code including the force field. As such, it fails to comply with the requirement of open and reproducible science and is thus harder to compare with other approaches. By contrast, a user friendly graphical interface for BAC, called ufBAC (see details below), has been developed in order to make it available to the widest range of users possible from academic to industrial and clinical. BAC has built-in ensemble-based simulation capabilities in order to ensure the accuracy and precision of reported results.

5.3. Distributed computing approaches

The significant computing resources required to deliver accurate binding predictions mean that we necessarily adopt a multitude of computing paradigms in order to perform investigations using BAC. To that end, we make use of large scale supercomputing class resources, as well as public cloud infrastructures including Azure, Amazon Web Services (AWS) and DNAnexus.

In the context of supercomputing class resources, we typically make use of RADICAL-Cybertools [150], developed by Rutgers University, a suite of tools that provide a common, consistent and scalable approach to high-performance and distributed computing. RADICAL-Cybertools consists of three fundamental components: (i) SAGA, an OGF community standard API for application-level jobs and data movement; (ii) RADICAL-Pilot (also known as BigJob), a tool that provides the ability to aggregate large number of tasks into a single-container job; (iii) EnsembleMD Toolkit, which builds upon RADICAL-Pilot as the execution layer, supports different patterns of ensemble-based computing, including replica exchange, workflows and simulation-analysis loops. These tools prove effective when wrapping complex workflows requiring very high core counts, and hence allow us to easily run replicas using the BAC Production component. RADICAL-Cybertools allows users to circumvent limitations put in place by supercomputing queuing systems, and run our applications in an efficient manner and ultra large scale including at the emerging exascale. A further feature of RADICAL-Cybertools allows adaptive applications to be created which alter the execution pattern based on the evolution of parameters in a defined set of simulations in order to promote computational efficiency.

RADICAL-Cybertools has been used to develop a version of BAC called high throughput or HT-BAC, designed to maximize simulation throughput when running on such high-performance computing platforms. Computational studies [8] have found that HT-BAC exhibits near linear weak and strong scaling when running both ESMACS and TIES BAC simulations, as shown in figure 2. Furthermore, the adaptive computing capabilities of RADICAL-Cybertools are employed by HT-BAC when running TIES calculations to automate runtime decisions based on partial simulation data and redistribute resources at runtime to support dynamically generated simulations.

Deploying BAC to cloud resources has required us to adopt technologies that minimize the differences between platforms. The rise of cloud computing has been accompanied by the development of so-called 'containerization' technologies, which allow a whole operating system to be virtualized [151]. Effectively this means that, using primarily the Docker [125] tool, it is possible to create a fixed, standard and portable environment for applications, by wrapping them in Docker containers. (On HPC platforms,

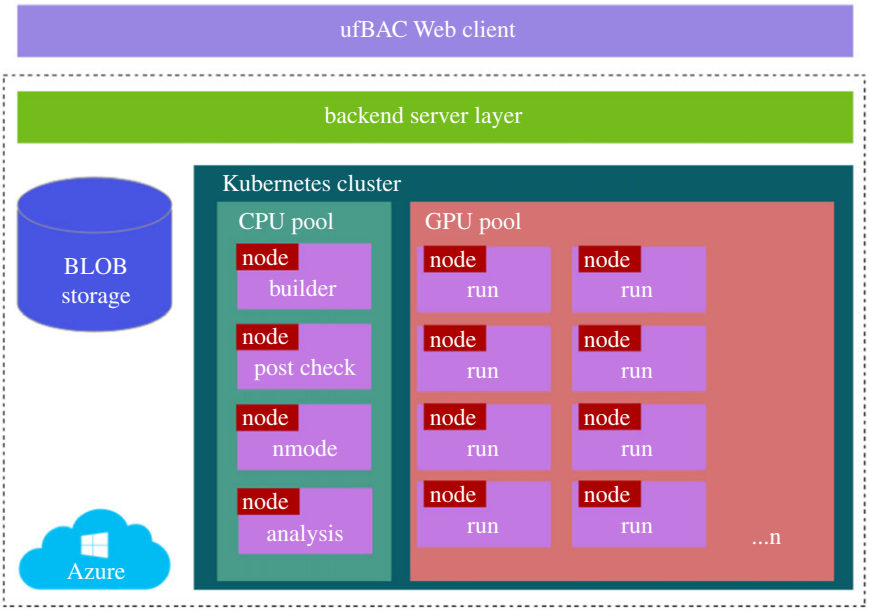

Figure 3. A deployment of the ufBAC application on the Azure cloud platform, using Kubernetes clusters and Docker containers. The main components of the application are encapsulated in Docker containers, the execution of which is managed by an application server layer. The user interacts with this layer via a Web portal.

which operate a different security model, Singularity is a preferred container technology—see below.) This means that the different components of the BAC application can each be embedded in their own container, including all of the dependencies required to run them. Each containerized BAC component is stored in a container registry, and can easily be deployed to any infrastructure that supports Docker (and/or Singularity).

This process is made easier using Kubernetes, an open source orchestration system for the automated deployment, scaling and management of the containerized application. Kubernetes allows us to define virtual clusters on a cloud, which can then be used to run containerized applications. These clusters can then be scaled up or down in order to execute the required number of replicas. In the context of the BAC application, it provides us with an infrastructure-independent platform on which to execute application components. Kubernetes is available as a managed service on Azure, AWS and Google Cloud Platform, meaning that the containerized BAC application can be easily ported between the leading cloud platforms, and indeed to any platform that supports Kubernetes and Docker. A base Docker image contains the core of each application component, and this is then used to build Docker images for specific cloud platforms that contain the necessary code to move data around that platform for example.

As well as facilitating platform-independent cloud deployments, the use of containerization is becoming increasingly widespread on supercomputing class resources, using either Docker or Singularity [139] (which is compatible with Docker containers). This means that containerized BAC components can easily be deployed to compatible supercomputing resources, without users having to have recourse to system administrators to install and optimize application dependencies.

But the ability to easily and reliably deploy components of the application on cloud and supercomputing resources is not sufficient to allow users to trivially perform complex investigations using the software. To remedy this, we have developed ufBAC [148], a Web portal interface to the BAC, which allows a user to build models of molecule-compound binding, and execute and analyse multi-replica MD simulations using the model. ufBAC enables BAC to be run via a Software as a Service model, hiding from the user the complexities of the command line tools used to build models, execute them and analyse the results. ufBAC is intended to plug in to a range of computational back ends, from HPC resources provided by academic national research facilities to commercial platforms such as Microsoft Azure or AWS.

The purpose of the ufBAC system, and the portal in particular, is to make the process of running complicated simulation workflows that rely heavily on HPC as simple as possible, improving usability by moving the user away from the command line towards a user friendly cloud style application. The ufBAC Web portal follows a conventional design. The left-hand side of the interface contains a menu bar that allows the user to access the various features of the application. The top bar of the website displays user notifications (for example that a set of simulations has finished running). The main content panel gives access to the features of the application and allows users to control running simulations, create and execute new models and analyse data.

A typical deployment of ufBAC, with BAC simulations performed on the Azure cloud using docker and Kubernetes, is outlined in figure 3. The application architecture comprises multiple layers. The first is the client layer, a user portal developed using Google Web Toolkit (GWT) [73]. The user interacts with the BAC system via their Web browser. The use of GWT provides a mechanism to develop high-performance, low overhead Web interfaces developed using Java which are compiled into separate JavaScript/HTML and Java byte code components, with the former running inside the user's browser (reducing server overheads) and interacting with the latter running inside a web application container such as Tomcat. It also means that new interfaces (designed for mobile devices, for example) can easily be constructed which make use of the common functionality provided by the server-side of the client.

The server layer comprises the server components of a set of RESTful Web services that allow the user to control the execution of the different BAC application components in

the cloud. Services support user login, state management and collaboration. It interfaces with the Kubernetes cluster used to run the BAC application components. This cluster comprises two to three compute pools, made up of differing types of cloud nodes (for example, a single core pool for building and analysing jobs, a GPU pool for production simulation, and a multicore pool to run normal mode calculations). In addition, a cloud blob store is used to store file-based output from the different components of the BAC application, and to pass data between components.

6. Application of free energy calculations

As we have emphasized, the most effective and reliable computational route to the reproducible ranking of the binding affinity of ligands to proteins can be achieved using ensemble methods. The endpoint approaches usually impose no significant restriction on the nature of the drug–protein systems that can be studied, although careful thought/attention needs to be given to the setting up of the models in many instances, such as the positions of metal ions and binding site water molecules [133], the parametrizations for the ligands, and so on. A study of box size dependence in simulations opens a debate [152–154], which indeed highlights the importance of setting up systems for simulation correctly and, more importantly, applying ensemble approaches to get statistically significant results.

We should point out that because of the flexibility of endpoint approaches, they are well suited to the early stage of drug discovery, so-called hit-to-lead. The approach does not provide accurate 'absolute' free energies, because of the nature of the approximations used, such as the implicit solvent models in MMPBSA-based approaches and the linear reaction assumption in LIE. However, the ensemble simulation based approaches can yield precise and reproducible, hence reliable, binding affinity ranking predictions. The alchemical approaches are in principle both accurate and precise in their domain of applicability [74]. The use of ensembles allows for the modelled systems to vary within a large phase space [21,25] and thereby describes a population of all relevant models which provide probabilistic predictions about the system behaviour.

The approach has been applied in different areas, including structural determination of proteins by combined experimental and computational information [155], and for structure predictions of ligand–protein complexes using docking approaches [156]. For binding free energy calculations, the ensemble approach is increasingly widely employed as the most effective way forward. Williams-Noonan *et al.* [157] have summarized some case studies using alchemical approaches. Here, we review some of the free energy applications, using endpoint and/or alchemical approaches, to relatively large datasets that closely mimic a real-world drug development setting, and to a few clinically approved drugs binding to sequence-dependent target proteins in a more forward-looking approach for personalized medicine. As the ensemble approaches are only now beginning to make their way into real-world problems, the applications reviewed below are not limited to ensemble-based methods. It needs to be emphasized once again here that, due to the random nature of MD trajectories, one-off simulations do not have any reliability; only ensemble simulations enable one to draw statistically significant conclusions [21].

In direct collaboration with various leading pharmaceutical companies, we have tested the ensemble free energy approaches in realistic pharmaceutical settings [59,60]. The calculations were performed, initially blind, to investigate the ability of our ESMACS and TIES methods to reproduce the experimentally measured trends which were released to us by the pharmaceutical companies after our computational predictions were made. Very good correlations were obtained from both of the methods. In addition to the binding free energy, structural, energetic and dynamic information at the atomistic level is forthcoming from the simulations, which cannot be obtained experimentally. Such information not only explains experimental observations, it sheds light on how to make modifications in the laboratory to improve the ligand binding and/or ligand selectivity [59,60].

Wang *et al.* [73] published a study with a large number of compounds binding to eight proteins. A total number of 330 relative binding free energies were calculated using single trajectory-based FEP+. While most of the published free energy studies focused primarily on retrospective predictions, the study [73] also included two prospective projects, where some of the compounds had been synthesized based on the alchemical free energy calculations. While the description of the prospective study was brief without revealing the structures of the compounds simulated and synthesized in this study, more details have been presented in a similar prospective study of GPCRs [158]. Based on the computational predictions, four novel compounds were synthesized and experimentally tested, showing that simulations correctly predicted the binding affinities for two of them. FEP+ was also used for the calculation of relative binding free energies of fragment-sized compounds using several pharmaceutically relevant targets with 96 fragments [159]. The studies demonstrate that such alchemical approaches have the potential to guide the synthesis of potent compounds, to impact fragment-based affinity optimization and to assist rational drug design projects.

Another less explored potential application of free energy prediction approaches is in the area of personalized medicine [82]. Due *inter alia* to the acquisition of drug resistance by individuals, it is often necessary to tailor the medication of a person according to his/her genetic configuration. In such cases, an accurate ranking of the available drugs based on their binding affinities when bound to different mutants of the target protein is the prerequisite. We studied the efficacy of two inhibitors to wild-type and mutant fibroblast growth factor receptor 1 (FGFR1) using ensemble simulations [75]. FGFR1 is a recognized therapeutic target in cancer. The binding affinities we predicted were confirmed by later-revealed biochemical measurements from our laboratory-based colleagues. The accuracy of the results displays the potential for the method to be applied in personalized medicine. Hauser *et al.* [160] recently published a study to predict how protein mutations modulate inhibitor affinities to Abl kinase. Classification of mutations as resistant or susceptible was predicted with a reasonable accuracy for eight FDA-approved drugs across 144 clinically identified mutations. Fowler *et al.* [161] demonstrated that ensemble alchemical approaches generated quantitatively accurate free energies, and were able to estimate the specificity and sensitivity of mutations in *Staphylococcus aureus* dihydrofolate reductase.

Point mutations in proteins can occur either inside ('local' mutants) or away from ('remote' mutants) the binding pocket. It is worth mentioning here that alchemical free

energy methods may not be able to correctly predict binding affinities in the case of remote mutants—the ones spatially a long way from the binding site—on typical wall clock time scales. Even using accelerated sampling techniques like REST2 cannot guarantee to improve the accuracy of such predictions [77] for reasons we alluded to earlier. Indeed an apparent underestimation from a recent study [78] calls attention to the need of further studies to validate the REST2 method in free energy calculations. A critical analysis of the application of alchemical free energy methods on protein mutations has been performed recently in our group [61]. There, we provide insights underpinning the impact of the gatekeeper mutation (a 'local' mutant) of FGFR3 on drug efficacy using ensemble approaches with the REST2 method. We focus on the UQ in these methods and, using that metric, we are able to compare the performance of different software and hardware for the calculation of the same free energy changes and show that, using ensemble-based methods, one can achieve reproducible results.

7. Conclusion

Drug–receptor binding is of key importance in determining drug efficacy and safety. The molecular determinants of binding affinity, compared with those of binding kinetics, are well understood. Binding free energy calculations are, therefore, expected to provide valuable contributions in real-world problems such as in rational drug design as a virtual screening and optimization tool, and in personalized medicine as a component of clinical decision support systems [82,162]. Numerous approaches and software tools have been developed for the purpose. Despite the applicability of the technology being well established, the methods have not thus far become standard virtual screening tools within the pharmaceutical industry, still less for decision support systems in clinical practice. The latter is a very new application of free energy calculations, being discussed only in recent years [18,26,82]. What aspects are limiting its applicability at present, and how can significant progress be made in the future?

While recent developments of the approaches have resulted in major improvements over what was available just a few years ago, there are still limitations in applying them within industry. These limitations include: the accuracy of the predictions, the challenge of handling truly diverse datasets, the general usability, as well as the computational power and financial cost required. While all of these limitations have been alleviated by the recent advances in software, middleware and hardware, novel approaches are required to further improve the accuracy of the predictions. Diverse datasets, including the incorporation of a variety of crystal structures, can succumb to careful preparation and analysis [40], including careful use of and standardization of the software, choice of force fields and assignment of partial charges, as well as selection of bound water molecules to include in the computations. It should be noted that often the sought correlation of the computed free energy is done against something not rigorously related to it, such as IC_{50}, and insufficient attention has been paid to the errors in those experimental measurements [14,51,73,163]. To date only a limited number of studies have been reported which compare free energy calculations from different MD codes and force fields due to technical difficulties of comparing with every code. Studies have shown that consistent results can be obtained across different MD engines [41,61,78]. Our recent investigations using three MD engines and two force fields show that the influences of force fields and MD codes on results are often quite small if ensemble methods are used as the basis for such comparisons [78].

In different scenarios, the cost–performance ratio can be in favour of cloud environment or on premise HPC facilities. While the former is usually preferred over the latter for small applications, traditional HPC facilities are what large scale tightly coupled calculations usually demand [164]. The equivalent instance hours required for MD simulations usually make the cloud applications prohibitive for many users, certainly in the academic community where investigators lack the supplies of cash required to meet such bills, but also in many companies; moreover, use of ensemble simulation substantially increases the costs. These factors make industrial users dither over committing to clouds when they could buy on premise HPC hardware. The current situation is that many users still prefer to own their own computers to avoid issues associated with the security of data, as well as cost. Nonetheless, the concepts of containerization and virtual machines are important advances brought about by the cloud computing paradigm which have entered mainstream HPC too in recent times [151]. In the long term, it is likely that both will be used, perhaps with bursting out to off-premises clouds when the workload cannot be handled internally. As mentioned above in the context of ML, reducing computational cost and time to solution is one of the motivations for invoking ML methods. However, UQ applied to ML predictions is in its infancy, being less mature than that we apply to MD simulation data.

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 by extending the length of simulations. Ensemble approaches are scalable, allowing thousands of binding affinities to be calculated per day, depending only on the computing resources available. Using ESMACS, for example, a single ligand–protein target can be assessed in 1–2 h using GPUs [61]. TIES is more computationally expensive, but predicts changes in free energy between pairs of structurally closely related ligand–protein systems. A binding free energy difference can be calculated in 2 h with modern codes running on accelerators, using HPC or cloud resource offering the latest GPU technology. Automated workflows are essential, which significantly increases the usability of the methods, while scale out to very large supercomputers makes it possible to deliver actionable predictions. Indeed, as evidenced by our Giant Workflow [165] on the entirety of Phase 1 and Phase 2 of SuperMUC, two linked supercomputers at the Leibniz Supercomputing Centre in Garching near Munich, with a combined total of about 245 000 cores where we achieved a sustained performance of about seven petaflops, it is now possible for us to produce high-quality binding affinity predictions for of the order of several hundred target proteins within a day or so with suitable computing resources. It is foreseeable that in the near future, rapid and accurate free energy prediction at high throughput will assist medicinal chemists in planning and directing compound synthesis in a routine manner.

Ensemble simulation-based free energy prediction approaches provide a route to predict relevant drug–protein binding affinities and hence are directly applicable in rational drug development and personalized medicine. They yield precise and reproducible, hence reliable, binding affinity ranking

predictions. They should provide a major boost to rational drug development in the pharmaceutical industry, and to personalized medicine in clinical practice. To be sure, there are still obstacles remaining, for example, in modelling charge-changing mutations and sampling relevant phase space when large conformational transitions occur. For mutations involving net charge changes, the long-range electrostatic interactions need to be incorporated properly [106]. Substantial conformational changes, which could be triggered by mutations involving charge changes or large size changes [61], are still a major issue for the convergence of free energy predictions. The double annihilation approach in ABFE calculations will almost always engender large conformational changes; the accuracy and convergence of the predictions need to be improved for the method to be used for a wider set of biologically interesting problems. The calculations are very computationally intensive, which is related to the requirement of sampling a representative conformational ensemble so as to get accurate and precise predictions. More efficient conformational search methods, which can access all of the energetically important conformational states, will significantly enhance the value of MD-based free energy calculations. With the improving theories and models, the increasing availability of automated tools and access to yet more powerful computing resources, binding free energy calculations are coming of age as a computational tool for the pharmaceutical industry and, in the longer term, to clinicians for drug selection in the context of personalized medicine.

Data accessibility. This article has no additional data.

Authors' contributions. All authors participated in the design of the study, contributed to the writing of the paper, gave final approval for publication and agreed to be held accountable for the work performed herein.

Competing interest. We declare we have no competing interests.

Funding. The authors would like to acknowledge the support of the MRC Medical Bioinformatics project (grant no. MR/L016311/1), the Qatar National Research Fund (grant no. 7-1083-1-191), the EU H2020 projects ComPat (grant no. 671564), CompBioMed (grant no. 675451), CompBioMed2 (grant no. 823712) and VECMA (grant no. 800925), NSF Award (award no. NSF 1713749) and funding from the UCL Provost. We made use of BlueWaters at the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, access to which was made available through the aforementioned NSF award; Titan and Summit at the Oak Ridge Leadership Computing Facility, supported by the Office of Science of the U.S. Department of Energy (DoE) under contract no. DE-AC05-00OR22725; SuperMUC at the Leibniz Supercomputing Centre in Garching, Germany; and cloud resources made available via DNAnexus and Microsoft Azure.

Acknowledgements. We thank a multitude of people for their assistance and helpful discussions, especially David W. Wright (UCL); Mike Kiernan and Sid Chaturvedi (Microsoft); Chai Fungtammasan, Brett Hannigan and Fiona Ford (DNAnexus); Sarah Skerratt, Kiyoyuki Omoto, Sharan K. Bagal and Veerabahu Shanmugasundaram (Pfizer); Ian Wall, Darren Green, Eric Manas, Alan Graves and Paul Bamborough (GlaxoSmithKline); Christophe Meyer, Herman van Vlijmen, Gary Tresadern and Laura Pérez-Benito (Janssen); and Ola Engkvist (AstraZeneca). We thank Dr Mateusz Bieniek (UCL) for providing access to his TIES simulation data.

References

1. Coveney PV, Dougherty ER, Highfield RR. 2016 Big data need big theory too. *Philos. Trans. A Math. Phys. Eng. Sci.* **374**, 20160153. (doi:10.1098/rsta.2016.0153)
2. Succi S, Coveney PV. 2019 Big data: the end of the scientific method? *Philos. Trans. A Math. Phys. Eng. Sci.* **377**, 20180145. (doi:10.1098/rsta.2018.0145)
3. 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.* **3**, 1900122. (doi:10.1002/adts.201900122)
4. DiMasi JA, Grabowski HG, Hansen RW. 2016 Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **47**, 20–33. (doi:10.1016/j.jhealeco.2016.01.012)
5. Peng RD. 2011 Reproducible research in computational science. *Science* **334**, 1226–1227. (doi:10.1126/science.1213847)
6. Springer Nature. 2020 Challenges in irreproducible research. See https://www.nature.com/collections/prbfkwmwvz/ (accessed 6 August 2020).
7. Hoekstra AG, Portegies Zwart S, Coveney PV. 2019 Multiscale modelling, simulation and computing: from the desktop to the exascale. *Philos. Trans. A Math. Phys. Eng. Sci.* **377**, 20180355. (doi:10.1098/rsta.2018.0355)
8. Dakka J, Turilli M, Wright DW, Zasada SJ, Balasubramanian V, Wan S, Coveney PV, Jha S. 2018 High-throughput binding affinity calculations at extreme scales. *BMC Bioinf.* **19**, 482. (doi:10.1186/s12859-018-2506-6)
9. Chong LT, Saglam AS, Zuckerman DM. 2017 Path-sampling strategies for simulating rare events in biomolecular systems. *Curr. Opin. Struct. Biol.* **43**, 88–94. (doi:10.1016/j.sbi.2016.11.019)
10. Groen D *et al*. 2019 Introducing VECMAtk: verification, validation and uncertainty quantification for multiscale and HPC simulations. In *Computational science – ICCS 2019* (eds JMF Rodrigues, PJS Cardoso, J Monteiro, R Lam, VV Krzhizhanovskaya, MH Lees, JJ Dongarra, PMA Sloot), pp. 479–492. Berlin, Germany: Springer International Publishing.
11. Foiles S, McDowell DL, Strachan A. 2019 Preface for focus issue on uncertainty quantification in materials modeling. *Model. Simul. Mater. Sci. Eng.* **27**, 080301. (doi:10.1088/1361-651x/ab46d6)
12. Thompson AL. 2019 Chemical crystallography: when are ' bad data' 'good data'? *Crystallogr. Rev.* **25**, 3–53. (doi:10.1080/0889311X.2019.1569643)
13. Coveney PV, Groen D, Hoekstra AG (eds). In preparation. Reliability and reproducibility in computational science: implementing verification, validation and uncertainty quantification *in silico*.
14. 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)
15. Deser C *et al*. 2020 Insights from Earth system model initial-condition large ensembles and future prospects. *Nat. Clim. Change* **10**, 277–286. (doi:10.1038/s41558-020-0731-2)
16. Palmer T. 2019 The ECMWF ensemble prediction system: looking back (more than) 25 years and projecting forward 25 years. *Q. J. R. Meteorol. Soc.* **145**, 12–24. (doi:10.1002/qj.3383)
17. Manos S, Zasada SJ, Coveney PV. 2008 Life or death decision-making: the medical case for large-scale, on-demand grid computing. *CTWatch Quarterly* **4**, 1–9.
18. Sadiq SK *et al*. 2008 Patient-specific simulation as a basis for clinical decision-making. *Philos. Trans. A Math. Phys. Eng. Sci.* **366**, 3199–3219. (doi:10.1098/rsta.2008.0100)
19. Balis B, Brzoza-Woch R, Bubak M, Kasztelnik M, Kwolek B, Nawrocki P, Nowakowski P, Szydlo T, Zielinski K. 2018 Holistic approach to management of IT infrastructure for environmental monitoring and decision support systems with urgent computing capabilities. *Future Gener. Comput. Syst.* **79**, 128–143. (doi:10.1016/j.future.2016.08.007)
20. Kovalchuk SV, Krotov E, Smirnov PA, Nasonov DA, Yakovlev AN. 2018 Distributed data-driven platform for urgent decision making in cardiological ambulance control. *Future Gener. Comput. Syst.* **79**, 144–154. (doi:10.1016/j.future.2016.09.017)
21. 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)

22. Rabier F, Klinker E, Courtier P, Hollingsworth A. 1996 Sensitivity of forecast errors to initial conditions. *Q. J. R. Meteorol. Soc.* **122**, 121–150. (doi:10.1002/qj.49712252906)
23. Yun-yu S, Mark AE, Cun-xin W, Fuhua H, Berendsen HJC, Gunsteren WFV. 1993 Can the stability of protein mutants be predicted by free energy calculations? *Protein Eng. Des. Sel.* **6**, 289–295. (doi:10.1093/protein/6.3.289)
24. 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)
25. 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)
26. 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)
27. van Gunsteren WF, Daura X, Hansen N, Mark AE, Oostenbrink C, Riniker S, Smith LJ. 2018 Validation of molecular simulation: an overview of issues. *Angew. Chem. Int. Ed. Engl.* **57**, 884–902. (doi:10.1002/anie.201702945)
28. Coveney PV, Highfield RR. 1991 *The arrow of time: the quest to solve science's greatest mystery*. London, UK: Flamingo.
29. Gallager RG. 2013 *Stochastic processes: theory for applications*. Cambridge, UK: Cambridge University Press.
30. 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)
31. 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)
32. Wan S, Sinclair RC, Coveney PV. 2020 Uncertainty quantification in classical molecular dynamics. (https://arxiv.org/abs/2006.07104)
33. 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)
34. Efron B, Tibshirani RJ. 1994 *An introduction to the bootstrap*. CRC press: FL, USA.
35. Li X, Wong W, Lamoureux EL, Wong TY. 2012 Are linear regression techniques appropriate for analysis when the dependent (outcome) variable Is not normally distributed? *Invest. Ophthalmol. Vis. Sci.* **53**, 3082–3083. (doi:10.1167/iovs.12-9967)
36. Frisch U. 1995 *Turbulence: the legacy of A. N. Kolmogorov*. Cambridge, UK: Cambridge University Press.
37. Adler M, Beroza P. 2013 Improved ligand binding energies derived from molecular dynamics: replicate sampling enhances the search of conformational space. *J. Chem. Inf. Model.* **53**, 2065–2072. (doi:10.1021/ci400285z)
38. Lawrenz M, Baron R, McCammon JA. 2009 Independent-trajectories thermodynamic-integration free-energy changes for biomolecular systems: determinants of H5N1 avian influenza virus neuraminidase inhibition by peramivir. *J. Chem. Theory Comput.* **5**, 1106–1116. (doi:10.1021/ct800559d)
39. Jiang W, Hodoscek M, Roux B. 2009 Computation of absolute hydration and binding free energy with free energy perturbation distributed replica-exchange molecular dynamics (FEP/REMD). *J. Chem. Theory Comput.* **5**, 2583–2588. (doi:10.1021/ct900223z)
40. Dakka J, Farkas-Pall K, Balasubramanian V, Turilli M, Wan S, Wright DW, Zasada S, Coveney PV, Jha S. 2018 Enabling trade-offs between accuracy and computational cost: adaptive algorithms to reduce time to clinical insight. In *2018 18th IEEE/ACM Int. Symp. on Cluster, Cloud and Grid Computing (CCGRID), Washington, DC, USA, 1-4 May 2018*, pp. 572–577. (doi:10.1109/CCGRID.2018.00005)
41. Loeffler HH, Bosisio S, Duarte Ramos Matos G, Suh D, Roux B, Mobley DL, Michel J. 2018 Reproducibility of free energy calculations across different molecular simulation software packages. *J. Chem. Theory Comput.* **14**, 5567–5582. (doi:10.1021/acs.jctc.8b00544)
42. Plesser HE. 2017 Reproducibility vs. replicability: a brief history of a confused terminology. *Front. Neuroinform.* **11**, 76. (doi:10.3389/fninf.2017.00076)
43. Ioannidis JPA. 2005 Why most published research findings are false. *PLoS Med.* **2**, 696–701. (doi:10.1371/journal.pmed.0020124)
44. Baker M. 2016 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454. (doi:10.1038/533452a)
45. Caves LS, 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)
46. Elofsson A, Nilsson L. 1993 How consistent are molecular-dynamics simulations: comparing structure and dynamics in reduced and oxidized *Escherichia coli* thioredoxin. *J. Mol. Biol.* **233**, 766–780. (doi:10.1006/jmbi.1993.1551)
47. Genheden S, Ryde U. 2010 How to obtain statistically converged MM/GBSA results. *J. Comput. Chem.* **31**, 837–846. (doi:10.1002/jcc.21366)
48. Frenkel D, Smit B. 2002 *Understanding molecular simulation: from algorithms to applications*, 2nd edn. San Diego, CA: Academic Press.
49. Genheden S, Ryde U. 2011 A comparison of different initialization protocols to obtain statistically independent molecular dynamics simulations. *J. Comput. Chem.* **32**, 187–195. (doi:10.1002/jcc.21546)
50. Manzoni F, Ryde U. 2018 Assessing the stability of free-energy perturbation calculations by performing variations in the method. *J. Comput. Aided Mol. Des.* **32**, 529–536. (doi:10.1007/s10822-018-0110-5)
51. 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)
52. Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, Shaw DE. 2012 Systematic validation of protein force fields against experimental data. *PLoS ONE* **7**, e32131. (doi:10.1371/journal.pone.0032131)
53. Lopes PEM, Guvench O, MacKerell Jr AD. 2014 Current status of protein force fields for molecular dynamics simulations. In *Molecular modeling of proteins* (ed. JM Walker), pp. 47–71. Totowa, NJ: Humana Press.
54. Uzun A, Leslin CM, Abyzov A, Ilyin V. 2007 Structure SNP (StSNP): a web server for mapping and modeling nsSNPs on protein structures with linkage to metabolic pathways. *Nucleic Acids Res.* **35**, W384–W392. (doi:10.1093/nar/gkm232)
55. Dellago C, Bolhuis PG. 2009 Transition path sampling and other advanced simulation techniques for rare events. In *Advanced computer simulation approaches for soft matter sciences III* (eds C Holm, K Kremer), pp. 167–233. Berlin, Germany: Springer.
56. Alder BJ, Wainwright TE. 1959 Studies in molecular dynamics. I. General method. *J. Chem. Phys.* **31**, 459–466. (doi:10.1063/1.1730376)
57. Bash PA, Singh UC, Brown FK, Langridge R, Kollman PA. 1987 Calculation of the relative change in binding free energy of a protein-inhibitor complex. *Science* **235**, 574. (doi:10.1126/science.3810157)
58. Genheden S, Ryde U. 2015 The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert. Opin. Drug. Discov.* **10**, 449–461. (doi:10.1517/17460441.2015.1032936)
59. 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)
60. 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)
61. 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)
62. Sinclair RC, Suter JL, Coveney PV. 2018 Graphene–graphene interactions: friction, superlubricity, and exfoliation. *Adv. Mater.* **30**, e1705791. (doi:10.1002/adma.201705791)
63. 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)

64. Abrams C, Bussi G. 2014 Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy* **16**, 163–199. (doi:10.3390/e16010163)
65. Bernardi RC, Melo MC.R, Schulten K. 2015 Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta* **1850**, 872–877. (doi:10.1016/j.bbagen.2014.10.019)
66. Valsson O, Parrinello M. 2014 Variational approach to enhanced sampling and free energy calculations. *Phys. Rev. Lett.* **113**, 090601. (doi:10.1103/PhysRevLett.113.090601)
67. Laio A, Gervasio FL. 2008 Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.* **71**, 126601. (doi:10.1088/0034-4885/71/12/126601)
68. 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)
69. 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)
70. Wang L, Berne BJ, Friesner RA. 2012 On achieving high accuracy and reliability in the calculation of relative protein–ligand binding affinities. *Proc. Natl Acad. Sci. USA* **109**, 1937–1942. (doi:10.1073/pnas.1114017109)
71. Paliwal H, Shirts MR. 2011 A benchmark test set for alchemical free energy transformations and its use to quantify error in common free energy methods. *J. Chem. Theory Comput.* **7**, 4115–4134. (doi:10.1021/ct2003995)
72. Shirts MR, Chodera JD. 2008 Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **129**, 124105. (doi:10.1063/1.2978177)
73. 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)
74. 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)
75. Schindler CEM *et al*. 2020 Large-scale assessment of binding free energy calculations in active drug discovery project. *J. Chem. Inf. Model.* (doi:10.1021/acs.jcim.0c00900)
76. Bunney TD *et al*. 2015 The effect of mutations on drug sensitivity and kinase activity of fibroblast growth factor receptors: a combined experimental and theoretical study. *EBioMedicine* **2**, 194–204. (doi:10.1016/j.ebiom.2015.02.009)
77. 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)
78. Wan S, Tresadern G, Pérez-Benito L, Vlijmen H, Coveney PV. 2019 Accuracy and precision of alchemical relative free energy predictions with and without replica-exchange. *Adv. Theory Simul.* **3**, 1900195. (doi:10.1002/adts.201900195)
79. Aldeghi M, Heifetz A, Bodkin MJ, Knapp S, Biggin PC. 2016 Accurate calculation of the absolute free energy of binding for drug molecules. *Chem. Sci.* **7**, 207–218. (doi:10.1039/c5sc02678d)
80. Babine RE, Bender SL. 1997 Molecular recognition of protein–ligand complexes: applications to drug design. *Chem. Rev.* **97**, 1359–1472. (doi:10.1021/cr960370z)
81. Sliwoski G, Kothiwale S, Meiler J, Lowe Jr EW. 2014 Computational methods in drug discovery. *Pharmacol. Rev.* **66**, 334–395. (doi:10.1124/pr.112.007336)
82. 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)
83. Mobley DL, Gilson MK. 2017 Predicting binding free energies: frontiers and benchmarks. *Annu. Rev. Biophys.* **46**, 531–558. (doi:10.1146/annurev-biophys-070816-033654)
84. Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR. 2008 Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* **153**(Suppl. 1), S7–S26. (doi:10.1038/sj.bjp.0707515)
85. Ballester PJ, Schreyer A, Blundell TL. 2014 Does a more precise chemical description of protein–ligand complexes lead to more accurate prediction of binding affinity? *J. Chem. Inf. Model.* **54**, 944–955. (doi:10.1021/ci500091r)
86. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. 2018 The rise of deep learning in drug discovery. *Drug Discov. Today* **23**, 1241–1250. (doi:10.1016/j.drudis.2018.01.039)
87. Perez A, Martinez-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)
88. Aqvist J, Luzhkov VB, Brandsdal BO. 2002 Ligand binding affinities from MD simulations. *Acc. Chem. Res.* **35**, 358–365. (doi:10.1021/ar010014p)
89. Kollman PA *et al*. 2000 Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **33**, 889–897. (doi:10.1021/ar000033j)
90. Jiménez-Luna J, Pérez-Benito L, Martínez-Rosell G, Sciabola S, Torella R, Tresadern G, De Fabritiis G. 2019 DeltaDelta neural networks for lead optimization of small molecule potency. *Chem. Sci.* **10**, 10 911–10 918. (doi:10.1039/C9SC04606B)
91. Sinitskiy AV, Pande VS. 2019 Physical machine learning outperforms 'human learning' in Quantum Chemistry. (https://arxiv.org/abs/1908.00971).
92. Aldeghi M, Gapsys V, de Groot BL. 2019 Predicting kinase inhibitor resistance: physics-based and data-driven approaches. *ACS Cent. Sci.* **5**, 1468–1474. (doi:10.1021/acscentsci.9b00590)
93. Calude CS, Longo G. 2017 The deluge of spurious correlations in big data. *Found. Sci.* **22**, 595–612. (doi:10.1007/s10699-016-9489-4)
94. Perez-Benito L, Casajuana-Martin N, Jimenez-Roses M, van Vlijmen H, Tresadern G. 2019 Predicting activity cliffs with free-energy perturbation. *J. Chem. Theory Comput.* **15**, 1884–1895. (doi:10.1021/acs.jctc.8b01290)
95. Romero R *et al*. 2019 Mechanism of glucocerebrosidase activation and dysfunction in Gaucher disease unraveled by molecular dynamics and deep learning. *Proc. Natl Acad. Sci. USA* **116**, 5086–5095. (doi:10.1073/pnas.1818411116)
96. Fox GC *et al*. 2019 Learning everywhere: pervasive machine learning for effective high-performance computation. In *2019 IEEE Int. Parallel and Distributed Processing Symp. Workshops (IPDPSW), Rio de Janeiro, Brazil, 20–24 May 2019*, pp. 422–429. (doi:10.1109/IPDPSW.2019.00081)
97. Rufa DA, Bruce Macdonald HE, Fass J, Wieder M, Grinaway PB, Roitberg AE, Isayev O, Chodera JD. 2020 Towards chemical accuracy for alchemical free energy calculations with hybrid physics-based machine learning/molecular mechanics potentials. *bioRxiv*. (doi:10.1101/2020.07.29.227959)
98. Homeyer N, Stoll F, Hillisch A, Gohlke H. 2014 Binding free energy calculations for lead optimization: assessment of their accuracy in an industrial drug design context. *J. Chem. Theory Comput.* **10**, 3331–3344. (doi:10.1021/ct5000296)
99. Swanson JM, Henchman RH, McCammon JA. 2004 Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys. J.* **86**, 67–74. (doi:10.1016/S0006-3495(04)74084-9)
100. 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)
101. 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**, 6017. (doi:10.1038/s41598-019-41758-1)
102. Wan S, Potterton A, Husseini FS, Wright DW, Heifetz A, Malawski M, Townsend-Nicholson A, Coveney PV. 2020 Hit-to-lead and lead optimization binding free energy calculations for G protein-coupled receptors. *Interface Focus* **10**, 20190128. (doi:10.1098/rsfs.2019.0128)
103. Gohlke H, Case DA. 2004 Converging free energy estimates: MM-PB(GB)SA studies on the protein-

protein complex Ras-Raf. *J. Comput. Chem.* **25**, 238–250. (doi:10.1002/jcc.10379)

104. Wan S, Coveney PV, Flower DR. 2005 Peptide recognition by the T cell receptor: comparison of binding free energies from thermodynamic integration, Poisson–Boltzmann and linear interaction energy approximations. *Phil. Trans. R. Soc. A* **363**, 2037–2053. (doi:10.1098/rsta.2005.1627)
105. Lin YL, Aleksandrov A, Simonson T, Roux B. 2014 An overview of electrostatic free energy computations for solutions and proteins. *J. Chem. Theory Comput.* **10**, 2690–2709. (doi:10.1021/ct500195p)
106. Chen W, Deng Y, Russell E, Wu Y, Abel R, Wang L. 2018 Accurate calculation of relative binding free energies between ligands with different net charges. *J. Chem. Theory Comput.* **14**, 6346–6358. (doi:10.1021/acs.jctc.8b00825)
107. Gapsys V, Pérez-Benito L, Aldeghi M, Seeliger D, van Vlijmen H, Tresadern G, de Groot BL. 2020 Large scale relative protein ligand binding affinities using non-equilibrium alchemy. *Chem. Sci.* **11**, 1140–1152. (doi:10.1039/C9SC03754C)
108. Fowler PW, Jha S, Coveney PV. 2005 Grid-based steered thermodynamic integration accelerates the calculation of binding free energies. *Phil. Trans. R. Soc. A* **363**, 1999–2015. (doi:10.1098/rsta.2005.1625)
109. Jha S, Coveney P, Harvey M. 2005 SPICE: simulated pore interactive computing environment. In *Proc. 2005 ACM/IEEE Conf. on Supercomputing, Seattle, WA, USA, 12–18 November 2005*. (doi:10.1109/SC.2005.65)
110. Boghosian B, Coveney P, Dong S, Finn L, Jha S, Karniadakis G, Karonis N. 2007 NEKTAR, SPICE and Vortonics: using federated grids for large scale scientific applications. *Clust. Comput.* **10**, 351–364. (doi:10.1007/s10586-007-0029-4)
111. Martin HSC, Jha S, Howorka S, Coveney PV. 2009 Determination of free energy profiles for the translocation of polynucleotides through α-hemolysin nanopores using non-equilibrium molecular dynamics simulations. *J. Chem. Theory Comput.* **5**, 2135–2148. (doi:10.1021/ct9000894)
112. Jorgensen WL, Buckner JK, Boudon S, Tirado-Rives J. 1988 Efficient computation of absolute free energies of binding by computer simulations. Application to the methane dimer in water. *J. Chem. Phys.* **89**, 3742–3746. (doi:10.1063/1.454895)
113. Samsudin F, Parker JL, Sansom MSP, Newstead S, Fowler PW. 2016 Accurate prediction of ligand affinities for a proton-dependent oligopeptide transporter. *Cell Chem. Biol.* **23**, 299–309. (doi:10.1016/j.chembiol.2015.11.015)
114. Shirts MR, Pande VS. 2001 Mathematical analysis of coupled parallel simulations. *Phys. Rev. Lett.* **86**, 4983–4987. (doi:10.1103/PhysRevLett.86.4983)
115. Zuckerman DM, Chong LT. 2017 Weighted ensemble simulation: review of methodology, applications, and software. *Annu. Rev. Biophys.* **46**, 43–57. (doi:10.1146/annurev-biophys-070816-033834)
116. Teo I, Mayne CG, Schulten K, Lelièvre T. 2016 Adaptive multilevel splitting method for molecular dynamics calculation of benzamidine-trypsin dissociation time. *J. Chem. Theory Comput.* **12**, 2983–2989. (doi:10.1021/acs.jctc.6b00277)
117. Potterton A, Husseini FS, Southey MWY, Bodkin MJ, Heifetz A, Coveney PV, Townsend-Nicholson A. 2019 Ensemble-based steered molecular dynamics predicts relative residence time of A2A receptor binders. *J. Chem. Theory Comput.* **15**, 3316–3330. (doi:10.1021/acs.jctc.8b01270)
118. Kokh DB *et al*. 2018 Estimation of drug-target residence times by tau-random acceleration molecular dynamics simulations. *J. Chem. Theory Comput.* **14**, 3859–3869. (doi:10.1021/acs.jctc.8b00230)
119. Altwaijry NA, Baron M, Wright DW, Coveney PV, Townsend-Nicholson A. 2017 An ensemble-based protocol for the computational prediction of helix–helix interactions in G protein-coupled receptors using coarse-grained molecular dynamics. *J. Chem. Theory Comput.* **13**, 2254–2270. (doi:10.1021/acs.jctc.6b01246)
120. Suter JL, Sinclair RC, Coveney PV. 2020 Principles governing control of aggregation and dispersion of graphene and graphene oxide in polymer melts. *Adv. Mater.* **32**, 2003213. (doi:10.1002/adma.202003213)
121. Pietrucci F. 2017 Strategies for the exploration of free energy landscapes: unity in diversity and challenges ahead. *Rev. Phys.* **2**, 32–45. (doi:10.1016/j.revip.2017.05.001)
122. Bernetti M, Masetti M, Rocchia W, Cavalli A. 2019 Kinetics of drug binding and residence time. *Annu. Rev. Phys. Chem.* **70**, 143–171. (doi:10.1146/annurev-physchem-042018-052340)
123. Buch I, Giorgino T, De Fabritiis G. 2011 Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations. *Proc. Natl Acad. Sci. USA* **108**, 10 184–10 189. (doi:10.1073/pnas.1103547108)
124. Ensign DL, Kasson PM, Pande VS. 2007 Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *J. Mol. Biol.* **374**, 806–816. (doi:10.1016/j.jmb.2007.09.069)
125. Cheng LS, Amaro RE, Xu D, Li WW, Arzberger PW, McCammon JA. 2008 Ensemble-based virtual screening reveals potential novel antiviral compounds for avian influenza neuraminidase. *J. Med. Chem.* **51**, 3878–3894. (doi:10.1021/jm8001197)
126. Wan S, Wright DW, Coveney PV. 2012 Mechanism of drug efficacy within the EGF receptor revealed by microsecond molecular dynamics simulation. *Mol. Cancer Ther.* **11**, 2394–2400. (doi:10.1158/1535-7163.MCT-12-0644-T)
127. Stjernschantz E, Oostenbrink C. 2010 Improved ligand–protein binding affinity predictions using multiple binding modes. *Biophys. J.* **98**, 2682–2691. (doi:10.1016/j.bpj.2010.02.034)
128. Smith M, Smith JC. 2020 Repurposing therapeutics for COVID-19: supercomputer-based docking to the SARS-CoV-2 viral spike protein and viral spike protein-human ACE2 interface. *ChemRxiv*. (doi:10.26434/chemrxiv.11871402.v4)
129. Duan Y, Kollman PA. 1998 Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740–744. (doi:10.1126/science.282.5389.740)
130. Georgoulia PS, Glykos NM. 2019 Molecular simulation of peptides coming of age: accurate prediction of folding, dynamics and structures. *Arch. Biochem. Biophys.* **664**, 76–88. (doi:10.1016/j.abb.2019.01.033)
131. Shaw DE *et al*. 2014 Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In *Proc. Int. Conf. for High Performance Computing, Networking, Storage and Analysis*, *New Orleans, LA, USA, 16–21 November 2014*, pp. 41–53. (doi:10.1109/SC.2014.9)
132. Ulmschneider JP, Ulmschneider MB. 2018 Molecular dynamics simulations are redefining our view of peptides interacting with biological membranes. *Acc. Chem. Res.* **51**, 1106–1116. (doi:10.1021/acs.accounts.7b00613)
133. Bodnarchuk MS. 2016 Water, water, everywhere… It's time to stop and think. *Drug Discov. Today* **21**, 1139–1146. (doi:10.1016/j.drudis.2016.05.009)
134. Karplus M. 2006 Spinach on the ceiling: a theoretical chemist's return to biology. *Annu. Rev. Biophys. Biomol. Struct.* **35**, 1–47. (doi:10.1146/annurev.biophys.33.110502.133350)
135. McCammon JA, Gelin BR, Karplus M. 1977 Dynamics of folded proteins. *Nature* **267**, 585–590. (doi:10.1038/267585a0)
136. Phillips JC *et al.* 2005 Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802. (doi:10.1002/jcc.20289)
137. Perilla JR, Goh BC, Cassidy CK, Liu B, Bernardi RC, Rudack T, Yu H, Wu Z, Schulten K. 2015 Molecular dynamics simulations of large macromolecular complexes. *Curr. Opin. Struct. Biol.* **31**, 64–74. (doi:10.1016/j.sbi.2015.03.007)
138. Hoekstra AG, Chopard B, Coster D, Portegies Zwart S, Coveney PV. 2019 Multiscale computing for science and engineering in the era of exascale performance. *Philos. Trans. A Math. Phys. Eng. Sci.* **377**, 20180144. (doi:10.1098/rsta.2018.0144)
139. Eastman P *et al*. 2017 OpenMM 7: rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659. (doi:10.1371/journal.pcbi.1005659)
140. Harvey MJ, Giupponi G, Fabritiis GD. 2009 ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.* **5**, 1632–1639. (doi:10.1021/ct9000685)
141. 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)
142. Homeyer N, Gohlke H. 2013 FEW: a workflow tool for free energy calculations of ligand binding. *J. Comput. Chem.* **34**, 965–973. (doi:10.1002/jcc.23218)

143. Wang K, Chodera JD, Yang Y, Shirts MR. 2013 Identifying ligand binding sites and poses using GPU-accelerated Hamiltonian replica exchange molecular dynamics. *J. Comput. Aided Mol. Des.* **27**, 989–1007. (doi:10.1007/s10822-013-9689-8)
144. Loeffler HH, Michel J, Woods C. 2015 FESetup: automating setup for alchemical free energy simulations. *J. Chem. Inf. Model.* **55**, 2485–2490. (doi:10.1021/acs.jcim.5b00368)
145. Gapsys V, Michielssens S, Seeliger D, de Groot BL. 2015 pmx: automated protein structure and topology generation for alchemical perturbations. *J. Comput. Chem.* **36**, 348–354. (doi:10.1002/jcc.23804)
146. Lundborg M, Lindahl E. 2015 Automatic GROMACS topology generation and comparisons of force fields for solvation free energy calculations. *J. Phys. Chem. B* **119**, 810–823. (doi:10.1021/jp505332p)
147. Kuhn M, Firth-Clark S, Tosco P, Mey ASJS, Mackey M, Michel J. 2020 Assessment of binding affinity via alchemical free-energy calculations. *J. Chem. Inf. Model.* **60**, 3120–3130. (doi:10.1021/acs.jcim.0c00165)
148. Zasada SJ, Wright DW, Coveney PV. 2020 Large-scale binding affinity calculations on commodity compute clouds. *Interface Focus* **10**, 20190133. (doi:10.1098/rsfs.2019.0133)
149. Liu S, Wu Y, Lin T, Abel R, Redmann JP, Summa CM, Jaber VR, Lim NM, Mobley DL. 2013 Lead optimization mapper: automating free energy calculations for lead optimization. *J. Comput. Aided Mol. Des* **27**, 755–770. (doi:10.1007/s10822-013-9678-y)
150. Ho TK. 1998 The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mech. Intell.* **20**, 832–844. (doi:10.1109/34.709601)
151. CompBioMed. 2020 Report on the use of commodity HPC infrastructures. See https://www.compbiomed.eu/wp-content/uploads/2019/02/D6.5-Report-on-the-Use-of-Commodity-HPC-Infrastructures.pdf (accessed 9 August 2020).
152. El Hage K, Hedin 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)
153. El Hage K, Hedin 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)
154. 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)
155. Bonomi M, Heller GT, Camilloni C, Vendruscolo M. 2017 Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **42**, 106–116. (doi:10.1016/j.sbi.2016.12.004)
156. Amaro RE, Baudry J, Chodera J, Demir O, 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)
157. Williams-Noonan BJ, Yuriev E, Chalmers DK. 2018 Free energy methods in drug design: prospects of 'Alchemical Perturbation' in medicinal chemistry. *J. Med. Chem.* **61**, 638–649. (doi:10.1021/acs.jmedchem.7b00681)
158. Lenselink EB *et al*. 2016 Predicting binding affinities for GPCR ligands using free-energy perturbation. *ACS Omega* **1**, 293–304. (doi:10.1021/acsomega.6b00086)
159. Steinbrecher TB, Dahlgren M, Cappel D, Lin T, Wang L, Krilov G, Abel R, Friesner R, Sherman W. 2015 Accurate binding free energy predictions in fragment optimization. *J. Chem. Inf. Model.* **55**, 2411–2420. (doi:10.1021/acs.jcim.5b00538)
160. Hauser K, Negron C, Albanese SK, Ray S, Steinbrecher T, Abel R, Chodera JD, Wang L. 2018 Predicting resistance of clinical Abl mutations to targeted kinase inhibitors using alchemical free-energy calculations. *Commun. Biol.* **1**, 70. (doi:10.1038/s42003-018-0075-x)
161. 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)
162. Wan S, Kumar D, Ilyin V, Homsi UA, Sher G, Knuth A, Coveney PV. 2020 From genome to personalised medicine: cancer treatment and discovery of novel variants in Qatar. Preprint.
163. Abel R, Wang L, Mobley DL, Friesner RA. 2017 A critical review of validation, blind testing, and real-world use of alchemical protein–ligand binding free energy calculations. *Curr. Top. Med. Chem.* **17**, 2577–2585. (doi:10.2174/1568026617666170414142131)
164. Netto MAS, Calheiros RN, Rodrigues ER, Cunha RLF, Buyya R. 2018 HPC cloud for scientific and business applications: taxonomy, vision, and research challenges. *ACM Comput. Surv.* **51**, 8. (doi:10.1145/3150224)
165. Simonson T, Archontis G, Karplus M. 2002 Free energy simulations come of age: protein-ligand recognition. *Acc. Chem. Res.* **35**, 430–437. (doi:10.1021/ar010030m)