Long timescale ensemble methods in molecular dynamics: Ligand-protein interactions and allostery in SARS-CoV-2 targets

Agastya P. Bhati[1‡], Art Hoti[1§‡], Andrew Potterton[2], Peter V. Coveney[*1,3,4]

Abstract

We subject a series of five protein-ligand systems which contain important SARS-CoV-2 targets - 3-chymotrypsin-like protease, papain-like protease and adenosine ribose phosphatase - to long-timescale and adaptive sampling molecular dynamics simulations. By performing ensembles of ten or twelve 10-microsecond simulations for each system, we accurately and reproducibly determine ligand binding sites, both crystallographically resolved and otherwise, thereby discovering binding sites that can be exploited for drug discovery. We also report robust, ensemble-based observation of conformational changes that occur at the main binding site of 3CLPro due to the presence of another ligand at an allosteric binding site. We investigate the reliability and accuracy of long-timescale trajectories. Due to the chaotic nature of molecular dynamics trajectories, individual trajectories do not allow for accurate or reproducible elucidation of macroscopic expectation values. Upon comparing the statistical distribution of protein-ligand contact frequencies for these ten/twelve 10-microsecond trajectories, we find that over 90% of trajectories have significantly different contact frequency distributions. Furthermore, using a direct binding free energy calculation protocol, we determine the ligand binding free energies for each of the identified sites using the long-timescale simulations. The free energies differ by 0.77 to 7.26 kcal/mol across individual trajectories depending on the binding site and the system. We show that although this is the standard way such quantities

[1] Centre for Computational Science, Department of Chemistry, University College London, London, United Kingdom. Tel: +44 (0)20 7679 4560; E-mail: p.v.coveney@ucl.ac.uk

[2] BenevolentAI, London, W1T 5HD, United Kingdom.

[3] Computational Science Laboratory, Institute for Informatics, Faculty of Science, University of Amsterdam, Amsterdam,The Netherlands.

[4] Advanced Research Computing Centre, University College London, London, United Kingdom.

§ Present address: Leiden Institute of Chemistry, Leiden University, Leiden, The Netherlands

‡ These authors contributed equally to this work

are currently reported at long-timescale, individual simulation does not yield reliable free energy. Ensembles of independent trajectories are necessary to overcome the aleatoric uncertainty in order to obtain statistically meaningful and reproducible results. Our findings here are generally applicable to all molecular dynamics based applications and not just confined to free energy methods used in this study. Finally, we compare the application of different free energy methods to these systems and discuss their advantages and disadvantages.

1 Introduction

There is an urgent need for drugs which target SARS-CoV-2, the pathogen responsible for the current coronavirus pandemic. In this regard, a concerted global effort has led to a rapid rise in the number of SARS-CoV-2 protein structures available in the Protein Data Bank (PDB), rendering the virus increasingly susceptible to rational, structure-based drug discovery. The typical timeline for the development of a single drug is 10-15 years, with an associated cost of $2 billion[1]. In the face of the global COVID-19 pandemic, it is clear that the average development timescale of up to 15 years is wholly inadequate. It is therefore of crucial humanitarian and societal importance to develop new *in silico* workflows that accelerate the rate and enhance the quality of lead drug molecule design. Workflows which tie both artificial intelligence (AI) and molecular dynamics (MD) based methods together are required as no single methodology can achieve both the required accuracy and speed[2–5]. Whilst AI based methods can rapidly sample significant regions of chemical space in a short time frame, MD based methods (which are significantly lower in throughput) are able to predict ligand binding free energies to much higher accuracy[2]. Furthermore, MD based methods have the potential to elucidate ligand binding kinetics and processes. The information derived from these simulations can be used to inform drug molecule optimisation for improved kinetic and thermodynamic binding properties. In turn, MD based methods form a crucial part of modern drug discovery workflows. In the present work, we investigate the application of molecular dynamics (MD) simulations to the robust and reproducible elucidation of ligand binding mechanisms, sites and interactions.

Molecular dynamics methods which aim to simulate the spontaneous process of protein-ligand binding have been in development for the last decade[6–14]. Over this period, significant advancements have been made due to increased access to high-performance computing resources (in particular GPU accelerated hardware), improvements in computational hardware[15,16], and developments in MD algorithms[17,18]. Thus far, work in the field has predominantly focused on determining the mechanism of ligand binding to crystallographically determined sites[6,7,9–12]. The idea behind these efforts is that by observing the spon-

taneous process of binding to these sites, key metastable states and associated protein ligand contacts can be identified. It is hoped that these interactions can then be modulated to optimise the kinetic and thermodynamic properties of drug binding[6,7,9–12]. Some of these studies have also led to the elucidation of non-experimentally determined sites, which may act as allosteric sites[7,8] for the modulation of protein activity.

A central problem that arises in these studies is that they utilise protocols which do not systematically account for the chaotic nature of molecular dynamics simulation[19]. The extreme sensitivity of such simulations to their initial conditions causes the many one-off results reported to be inherently non-reproducible[20–23]. Addressing this issue forms the central focus of this work. The question which we address is whether it is possible to develop reliable methods that can accurately and reproducibly identify the full range of binding sites and binding modes that are accessible to a ligand. Such a method will permit us to go beyond what is essentially anecdotal evidence, and to report that findings are statistically reliable and of scientific value. We would like to remind readers that it is common to work with fixed epistemic parameters in molecular dynamics. Although a full uncertainty quantification analysis would require one to investigate their role in determining the uncertainties in quantities of interest, we have previously shown that the aleatoric uncertainty in MD simulations typically overwhelms that from the epistemic sources, and that the latter's uncertainty is damped in the output quantities of interest[24]. Therefore, our focus in this work is only on the aleatoric uncertainty.

In general, spontaneous ligand binding methods work by initiating a molecular dynamics protein-ligand system from a configuration where the ligand is placed at some distance from the surface of the protein. During the simulation, the ligand explores the surface of the protein and binds to potentially druggable sites which may be orthosteric, allosteric or even cryptic in nature[8]. By analysing the trajectories using methods such as Markov state model (MSM) analysis[25,26], thermodynamic and kinetic observables which are of key importance to the process of drug discovery can be extracted from the data. These include binding free energies[27], dissociation constants (K_d)[27] and on and off rates of binding (K_{on} and K_{off})[28,29].

When conducting these studies, the question arises as to whether the trajectory has sufficiently sampled phase space such that the probability distribution of the trajectory has converged to the equilibrium probability distribution of the protein-ligand system. Only this distribution would allow the true expectation values of the observables to be obtained[19]. To sample the phase space, one of two distinct approaches is usually followed. In the first, which we term the "long-timescale" regime, authors report several microsecond timescale simulations[7–9,27,30] and from these, compute the observables of

interest. These observables include 3-dimensional ligand occupancy maps, ligand binding free energies, along with ligand association mechanisms and pathways. In the second regime, termed "adaptive-sampling"[11–13,29,31,32], many simulations of shorter timescale are executed, and new simulations are adaptively initiated from specific simulation snapshots in order to "more thoroughly" explore regions of phase space that are of interest. Incidentally, many studies from the second regime report aggregate simulation times that fall in the microsecond timescale; this is misleading as performing a single simulation of that duration is not the same as we will discuss in detail in the current study. We note that some studies also combine the two techniques, using adaptive sampling to initiate new, "short" simulations from long-timescale simulations that are stuck within non-productive kinetic traps[28]. Generally, this approach is taken in order to converge transition probabilities between metastable states that are identified during Markov state modelling[33,34].

Current practices in the field, however, do not *systematically* consider the aleatoric uncertainty associated to microsecond-timescale simulations. Molecular dynamics trajectories are intrinsically chaotic in nature, meaning that they exhibit extreme sensitivity to initial conditions[19–24,35–38]. This causes results derived from individual simulations to be non-reproducible. Indeed, within accessible timescales, the oft-made assumption of ergodicity according to which an ensemble average may be replaced by a time-average, does not hold[19].

We would like to point out here that there are several accelerated sampling protocols that involve performing "ensembles" of simulations. These include methods that do not employ any external force or heating, and just enhance sampling by performing multiple independent MD simulations concurrently with different starting conditions. Examples include ensemble dynamics[39–45], Markov state model (MSM)[25,26,33,34,43,45–48], weighted ensemble (WE)[49–58] and multilevel splitting (MS)[59–63] methods. Although these methods involve performing "replicas" and generating "ensembles", the fundamental question is whether we get the same answer (within error bars) on repeating the entire protocol using one of the above methods. Given that the dynamics is chaotic, it is expected that this is not the case and ensembles must be used as each execution of such a protocol would have a different initial condition[19,24]. One example are replica exchange methods[64,65] that also involve performing multiple MD simulations in parallel (so called "replicas"). We have shown in previous work that on repeating a replica exchange calculation multiple times, we indeed observe variation in the outcome and hence it is necessary to perform ensembles of the entire protocol (which itself contains "replicas") to perform a systematic uncertainty quantification (UQ)[21,66]. Similar studies are required for other methods involving "ensembles" in order to properly assess UQ in those cases.

The purpose of the present paper is to systematically assess the distribution of properties obtained in long-timescale simulations. By performing ensembles of 10 to 12 ten-microsecond unbiased simulations, we are able to evaluate the utility of running individual long-timescale simulations, investigate their reproducibility and compare the results obtained from them to an "adaptive-sampling" scheme which consists of 9 microseconds of *aggregate* simulation time. This is of interest as the wall time required to execute the long-timescale runs is significantly longer than the wall time required for the entire adaptive sampling protocol (differences are on the order of weeks).

In our study, we apply these statistically robust techniques to three crucial SARS-CoV-2 drug targets: adenosine ribose phosphatase (ADRP)[67], papain-like protease (PLpro)[68] and 3-chymotrypsin-like protease (3CLpro)[69]. Each of these are globular, non-structural proteins encoded by SARS-CoV-2 which play key roles in the lifecycle of the virus and serve as important potential targets for SARS-CoV-2. Our findings here shed light on potentially druggable sites on the surface of the proteins, elucidate relative binding free energies between each of the sites, and demonstrate binding mechanisms which may explicitly inform future efforts in SARS-CoV-2 drug discovery. We also compare different free energy protocols and discuss the applicability of each in different scenarios. In addition to these methodological developments, we report new scientific findings on the allosteric effects observed in 3CLPro. We discover conformational changes occurring at the active site of 3CLPro caused by the binding of a ligand at an experimentally known allosteric binding site. We demonstrate how these changes affect the binding of ligands at the main (active) site by distorting the binding pose. We also show how one-off simulations can easily lead us to draw false conclusions. Only ensemble simulations can provide statistically robust results.

2 Theory

The present study approaches the subject of spontaneous protein-ligand binding simulations from the perspective of chaos theory and uncertainty quantification. In this section, we describe how the chaotic nature of molecular dynamics simulations causes individual simulations to be non-reproducible, no matter their length. We also describe how running ensembles of simulations remedies this by allowing for expectation values to be subjected to rigorous uncertainty quantification and convergence analysis. By presenting this theory, we strive to make clear that running an ensemble of simulations that sum to a certain time is not equivalent to simply running a single simulation of the same aggregate time. The novel point which we explicitly demonstrate is that, contrary to the current consensus, the level of certainty of a simulation derived expectation value does not increase with simulation time. This

necessitates the use of ensembles when reporting macroscopic expectation values for all long-timescale simulations. Not only is such a prescription required by the tenets of statistical mechanics, it is also essential in order to quantify the uncertainty of the calculated properties.

2.1 MD simulation and equilibrium

In statistical mechanics, the value of an observable (G) of a dynamical system is derived by calculating the expectation value of the observable $\langle G \rangle_t$ over the trajectory that the dynamical system takes through phase space

$$\langle G \rangle_t = \int G(x)\rho_t(x)d\mu \tag{1}$$

The ergodic theorem, often used to justify the accuracy of "long-timescale" molecular dynamics simulations, states that in the long-time limit, the time average of a dynamical observable will approach its ensemble average. Namely,

$$\lim_{t\to\infty} \langle G \rangle_t = \langle G \rangle_{eq} = \lim_{t\to\infty} \int G(x)\rho_t(x)d\mu = \int G(x)\rho_{eq}(x)d\mu \tag{2}$$

where ρ_t and ρ_{eq} are the 6N+1 dimensional time dependent and equilibrium probability distributions of the dynamical system.This implies that ρ_t has asymptotically approached ρ_{eq} (where the evolution of ρ_t is determined by the Liouville equation[19]).

Problematically for those working in the field of MD simulation, this assumption only holds true for timescales that are on the order of a Poincaré recurrence time, which is longer than the age of the universe[23]. Therefore, because for any realistically obtainable single trajectory of a dynamical system ρ_t does not asymptotically approach ρ_{eq}, the value of a given observable obtained from an individual molecular dynamics trajectory cannot be equated to the true value of the observable that would arise if phase space were ergodically sampled. Furthermore, the equality also requires that the dynamical system is mixing. In the ergodic hierarchy, mixing is a stronger property than ergodicity, and is dependent on the system being chaotic[19].

2.2 Uncertainty quantification

Uncertainty quantification (UQ) is a field of endeavour that aims to analyse the interplay between simulation inputs and outputs for the purpose of determining the uncertainty associated to obtained results[23,70]. In the present study, we are particularly interested in quantifying the aleatoric output uncertainty that is controlled by the initial random velocity seed. A series of our studies have presented

robust evidence that simulation outcomes are strongly controlled by the initial random seed, and that averaging over a set of simulations all starting with different random seeds consistently reduces the uncertainty of obtained results[24,35]. Indeed, this aleatoric uncertainty completely dominates the epistemic uncertainty arising from the way in which the model is parameterised and set up[24]. A crucial feature of performing ensembles of simulations, which allows us to conduct uncertainty quantification, is that the *distribution* of properties of interest can be obtained. Here, we apply UQ to multiple properties of spontaneous ligand-protein binding simulations, namely, the protein-ligand residue contact frequency distribution and the computed binding free energy of the ligand with a protein target.

2.3 SARS-CoV-2 protein-ligand systems

Three important SARS-CoV-2 targets form the focus of this work: 3CLpro, PLpro and ADRP. 3-chymotrypsin-like protease (3CLpro, also known as the main protease, or non-structural protein 5 (nsp5)), and papain-like-protease (PLpro, the protease domain of nsp3) are both proteolytic enzymes of SARS-CoV-2 which are responsible for cleaving the viral poly-protein chain (encoded by SARS-CoV-2 RNA) into non-structural proteins that are required for the process of viral replication[68,69]. Adenosine ribose phosphatase (ADRP) is a domain of Nsp3 that is capable of interfering with the host immune response by removing ADP-ribose from ADP ribosylated proteins and RNA[67]. Thus, each of these protein targets are of considerable interest for SARS-CoV-2 drug design.

In a recent study by our group[38], 14 compounds of interest, each of which bind to one of three sites (the substrate binding site, allosteric site I, and allosteric site II) on the surface of 3CLpro, were identified from a previously conducted high-throughput crystallographic screen of repurposed drug molecules[71]. Based on the results derived in that study, we selected 3 ligands of interest, MUT056399 (RQN), AT7519 (LZE) and pelitinib (93J) for the current study covering all three binding sites and a wide range of EC50 values (Table 1). Furthermore, by building the system containing both RQN (which binds to the substrate binding site)[71] and LZE (which binds to allosteric site 2)[71], we aim to capture whether or not the binding of RQN is affected by the binding of LZE, and if so then determine the allosteric mechanism involved in the process. For the PLpro system we decided to focus on the ligand GRL0617 as it showed strong antiviral activity using NMR data and a promising value of EC50[68]. Tofacitinib, which is a FDA approved pharmaceutical that is used to treat rheumatoid arthritis and ulcerative colitis[72,73] was chosen as the ligand for the ADRP system.

Table 1: Protein targets and their corresponding ligands. [a]PDBe ligand codes. [b]6W02 is the structure of ADRP bound to ADPR, not to tofacitinib.

Target name	Compound name	PDBe[a]	Exp. Binding site	PDB ID	EC50 (μM)
ADRP	Tofacitinib	MI1	N/A	6W02[b]	N/A
PL-Pro	GRL-0617	TTT	USP	7CJM	21.00
3CL-Pro	MUT056399	RQN	SB	7AP6	38.24
	Pelitinib	93J	AS I	7AXM	1.25
	AT7519	LZE	AS II	7AGA	25.16

3 Methods

We use ensembles of replica simulations (which here differ only in their initial particle velocities, drawn randomly from a Maxwell-Boltzmann distribution) in order to converge the statistics of the observable of interest. Whilst previous studies by our group have investigated the necessity of ensembles for accurate and precise ligand binding free energy calculations[20,22,74–76], here we aim to demonstrate that ensembles of MD simulations are equally essential for the accurate determination of ligand binding sites and ligand-protein interaction mechanisms. To do this, we conduct a thorough comparative analysis of two alternative ensemble protocols: the long-timescale protocol, and the splitting protocol (an adaptive sampling method). These protocols are applied with a key goal in mind: to elucidate novel ligand binding sites and mechanisms for the three aforementioned proteins that are essential to the life cycle of SARS-CoV-2: ADRP, 3CL-Pro, and PL-Pro. We also employ different free energy protocols in order to determine the pros and cons of each method and discuss their domains of applicability and limitations.

3.1 Protein-ligand systems

All three protein systems were selected due to their key-role in the life-cycle of the SARS-CoV-2 virus (as discussed in §2). The protein structures were initially sourced from PDBs (see Table 1). All mutations in the initial crystallographic structures were back-mutated using the "Rotamers" tool in UCSF Chimera[77,78]. Following this, ligands and other unwanted molecules were removed from the structures. The 3-dimensional conformers of the selected ligands were sourced from PubChem (`https://pubchem.ncbi.nlm.nih.gov`) and inserted into the system. Five protein-ligand systems were built in total. All systems are detailed in Table 1 and each of the proteins and ligands are shown visually in Figure 1. For each of our protocols and systems, the ligand was initially placed 20Å away from the surface of the protein. A distance of 20Å was chosen to minimise sampling bias that would arise from the initial position of the ligand due to long-range protein-ligand interactions. Thus, if the ligand was initially

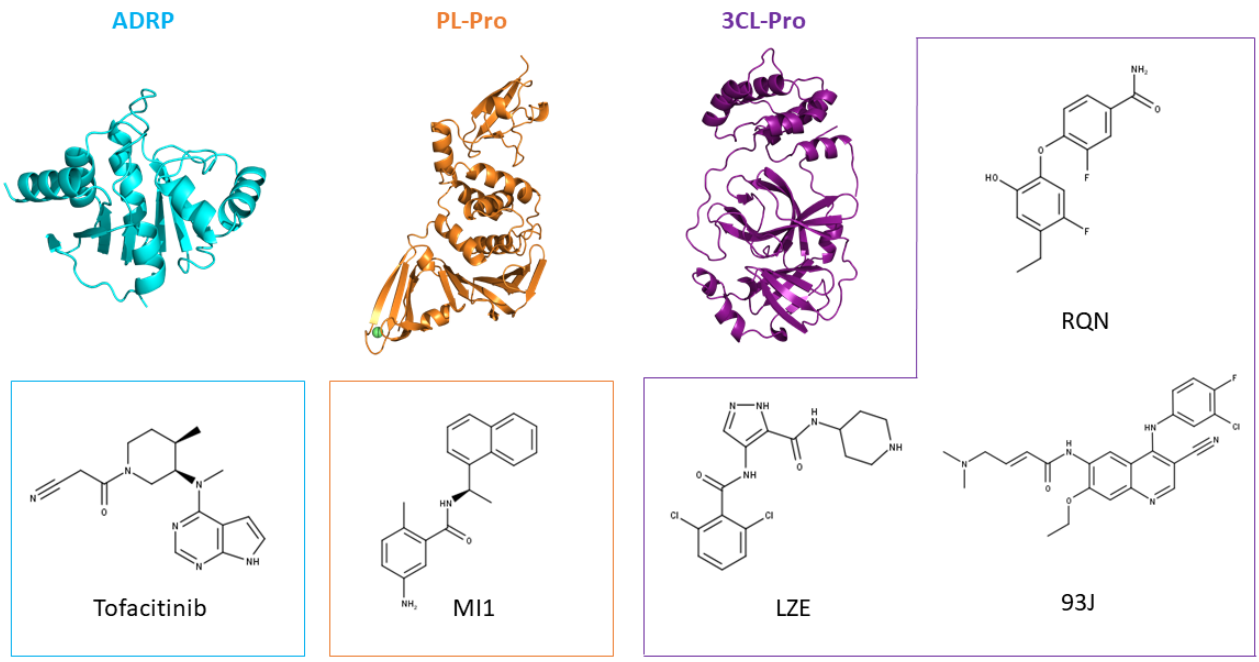

Figure 1: Structures of protein targets and corresponding ligands. ADRP (PDBID: 6W02) is shown in cyan, PLpro (PDBID: 7CJM) in orange and 3CLpro (PDBID's: 7AP6, 7AXM, 7AGA) in purple.

placed 3Å from the binding site, it would immediately form interactions with the protein in that region and therefore most likely bind to that site. By distancing the ligand we ensure that it stochastically diffuses around the protein before establishing its initial contact. Furthermore, the choice of separating the ligand and the protein by a distance of 20Å is compatible with standard practice in the field, which is to distance the ligand between 20 and 30Å away from the surface of the protein[6,7,10,11]. Following this, each system was solvated using the TIP3P water model and charge-neutralised by inserting sodium or chloride ions[79].

3.2 Simulations

In this sub-section, we describe the two simulation approaches which we directly compare within this study: The long-timescale and splitting protocols.

3.2.1 Long timescale protocol

In the long-timescale protocol (Fig. 2B), we perform ten or twelve replica simulations of 10 microseconds each. Each simulation is initiated from a common configuration in which the ligand is placed 20 Ångstroms from the surface of the protein. A simulation length of 10 microseconds is chosen on the basis

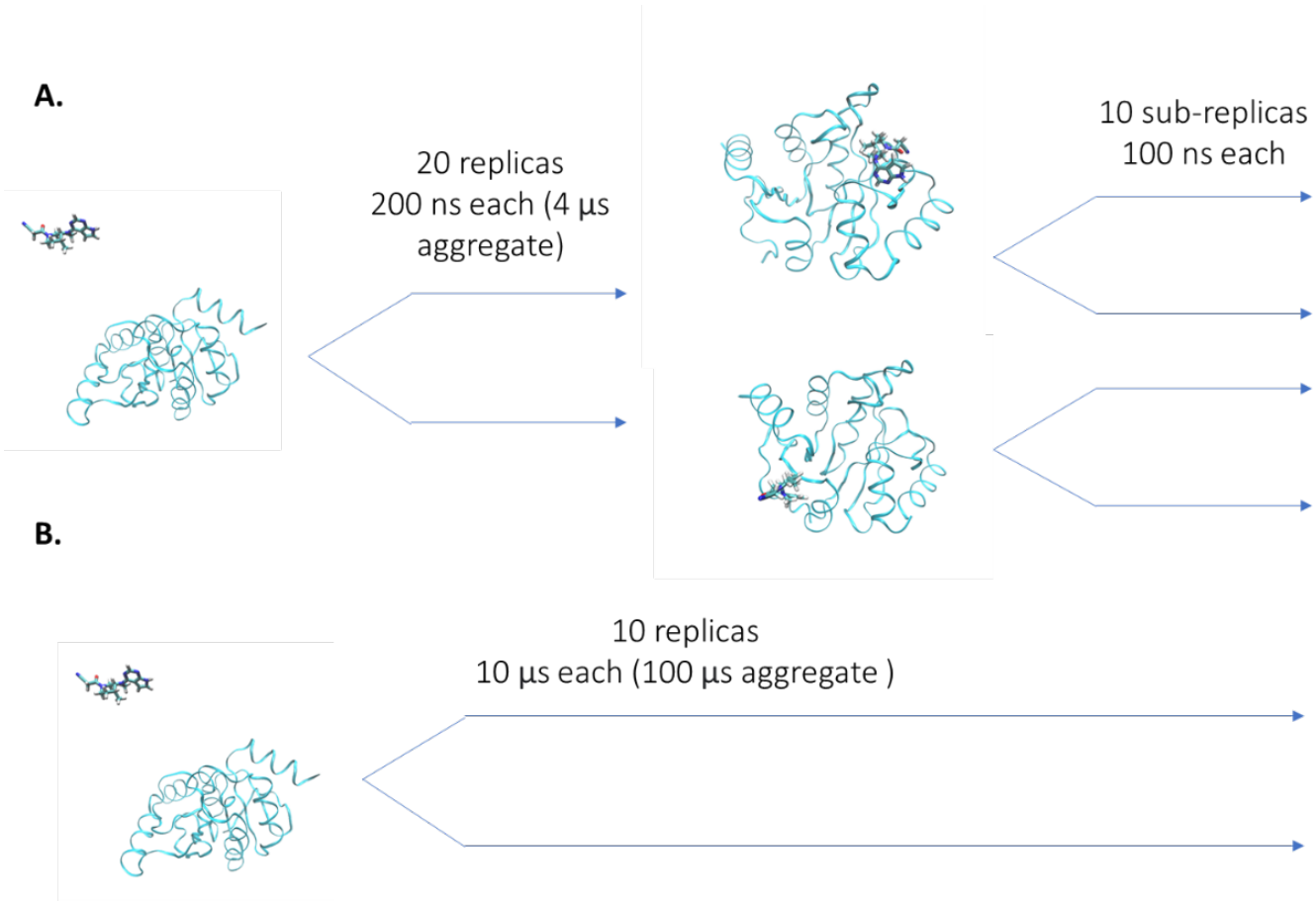

Figure 2: Schematic depiction of the "splitting" and "long-timescale" protocols. (A) Within the splitting protocol, ensembles of 20 replica trajectories of 200 nanoseconds are initiated from a common configuration in which the ligand is placed 20Å from the surface of the protein. For each replica, the initial particle velocities are drawn randomly from a Maxwell-Boltzmann distribution. Trajectories are analysed using RMSD heatmaps, and those with the most stable poses in the final frame are chosen as configurations from which to initiate new sets of replicas, which we term "subreplicas". Each set of "subreplicas" contains 10 subreplicas of 100 nanoseconds each. (B) During the long-timescale protocol, 10 or 12 replicas of 10 microseconds each are initiated from a common configuration in which the ligand is placed 20Å from the surface of the protein. For each replica, the initial particle velocities are drawn randomly from a Maxwell-Boltzmann distribution.

that it is on the order of simulation times (microseconds to tens of microseconds) that have been utilised in multiple previous studies to derive information on the nature of ligand-protein interactions[8,9,27] This protocol allows us to address two crucial aims within our study. First, we intend to identify whether a single 10 microsecond run can reliably reproduce the full range of binding sites and binding modes sampled by the aggregate of the "splitting" protocol (which has a length of 9 microseconds). And second, we aim to demonstrate the variability between the 10 microsecond members of the ensemble in order to examine whether a single "long-timescale" (ten microsecond) trajectory is capable of generating reproducible and therefore reliable results. Indeed, as we show, each 10-microsecond run exhibits different statistics due to the chaotic nature of MD trajectories. A few recent papers from the D. E. Shaw group implicitly recognise such variability in MD simulations at the microsecond timescale[80,81]. However, this has not been studied systematically hitherto, nor has the importance of ensembles of simulations at long-timescale been discussed in the literature as we do in this study.

3.2.2 Splitting protocol

During the splitting protocol (Fig. 2A), 20 replica trajectories of 200 nanoseconds each are initiated from a common configuration in which the ligand is placed 20 Ångstroms from the surface of the protein. The initial particle velocities of each ensemble member are drawn randomly from a Maxwell-Boltzmann distribution. Trajectories are analysed using RMSD heatmaps, and the 5 replicas with the most kinetically stable poses in the final frame are chosen as configurations from which to initiate new sets of replicas, which we term "subreplicas". We quantify "kinetic stability" by computing the ligand RMSD relative to the final frame of the simulation and select the five replicas which have an RMSD of < 5 Ångstroms for the longest duration of time relative to the final frame. Each set of "subreplicas" contains 10 subreplicas of 100 nanoseconds each. The aggregate simulation time across the length of this protocol is 9 microseconds. Within the protocol, 200 nanoseconds and 100 nanoseconds were chosen as the simulation times as these are representative of the simulation timescale executed by those who have utilized ensemble based adaptive sampling protocols. Examples of this include the seminal study in the field by Butch *et al.* where 495 trajectories of 100 nanoseconds each were executed[6], amongst other papers which run on similar timescales[10,11]. The purpose of the splitting method is to explore and identify as many binding sites as are feasible to which the ligand of interest may bind on the protein, whilst reducing the amount of wall time required to do so.

3.2.3 Simulation details

NAMD 2[79,82] and OpenMM[83] were used to run our simulations. All splitting protocol simulations were executed on Scafell Pike (`hartree.stfc.ac.uk`) whose compute nodes are comprised of Bullsequana X1000 (Intel Xeon processors and NVIDIA Tesla V100 accelerators). Long-timescale runs for ADRP were also executed on Scafell Pike using OpenMM. All other long timescale simulations were executed using OpenMM on Summit (`https://www.olcf.ornl.gov/summit`) where compute nodes consist IBM Power9 processors and NVIDIA Tesla V100 accelerators. Force fields and modifiable simulation parameters were kept constant across MD engines and HPC platforms. All ligands (Table 1 and Fig. 1) were parameterised in AmberTools using AM1-BCC charge assignments. The Amber FF14SB force field was used to parameterise the protein, and TIP3P water molecules were used to solvate the system. During equilibration, we conducted 1000 steps of energy minimisation, and then in the NVT ensemble, applied harmonic constraints to protein and ligand atoms, whilst heating the system from 60K to 310K (an increase of 1K every 2ps). We then ran in the NVT ensemble at 310K for 300ps with no constraints. Following this, we performed equilibration in the NPT ensemble, using a Monte-Carlo barostat with a pressure of 1.01325 bar and frequency of 50fs. We reduced the strength of all harmonic constraints by a half every 0.1 ps, 10 times. Subsequently, constraints were set to 0. Finally, the system was equilibrated without constraints at 310K in the NPT ensemble for 1ns. For all production and equilibration simulations, a Langevin thermostat was employed with a 2fs timestep together with a friction coefficient of 1/picosecond to simulate the dynamics of the system.

3.3 Ligand-protein contact frequency analysis

Ligand - protein residue contact frequencies are computed using a series of custom python scripts. The original scripts were written for the "getcontacts" tool by Dror et al[84]. A contact between the ligand and the protein is defined as a van der Waals interaction, where the distance ($|AB|$) between two non-hydrogen atoms, A (belonging to the ligand) and B (belonging to the protein), satisfies the equation: $|AB| < R_{vdW}(A) + R_{vdW}(B) + 0.5$, where R_{vdW} is the van der Waals radius of the atom.

Upon computing the percentage of frames in which contacts are formed between the ligand and each protein residue for all of our trajectories, we obtain a two-dimensional matrix containing $m \times n$ elements where m is the number of trajectories executed and n is the number of residues in the protein. An element (m, n) of the matrix therefore corresponds to the contact frequency of the ligand with residue n in trajectory m. All ligand-residue contact frequency distributions are computed from these

matrices using Python. These distributions provide meaningful and easily interpretable low dimensional representations of phase space sampling.

3.4 Binding free energy calculations

To determine the relative binding free energy of a specific ligand for each of its identified binding sites, we use two protocols: ESMACS[22,74,75], and the so-called "direct" binding free energy calculation method[27,85]. By running the direct protocol, we also derive insights into the reproducibility of expectation values that are computed from "converged" simulations that are multiple microseconds in length.

3.4.1 Enhanced sampling of molecular dynamics with approximation of continuum solvent (ESMACS)

Enhanced sampling of molecular dynamics with approximation of continuum solvent (ESMACS) calculations are fundamentally based on the Molecular Mechanics Poisson-Boltzmann/ Generalised-Born Surface Area (MMPB/GBSA) binding free energy calculation method[86]. MMPB/GBSA calculations were conducted using AmberTools 20[87]. For all MMPB/GBSA calculations, the 1-traj protocol was used, allowing the MMPB/GBSA ligand binding free energy ($\Delta G_{MMPB/GBSA}$) to be calculated from a single trajectory of the protein-ligand complex. Within the 1-traj protocol $\Delta G_{MMPB/GBSA}$ is computed using the equation:

$$\Delta G_{MMPB/GBSA} = \langle G_{PL} - G_P - G_L \rangle_{PL},$$

where G_{PL}, G_P and G_L correspond to the free energy contributions of the complex, protein, and ligand respectively. Angular brackets denote that $\Delta G_{MMPB/GBSA}$ is computed as the average over all input snapshots, while the subscript "PL" denotes that the snapshots are taken from a single simulation of the protein-ligand complex. G_{PL}, G_P and G_L are calculated using the following equation:

$$G = E_{bnd} + E_{ele} + E_{vdW} + G_{Pol} + G_{np},$$

where E_{bnd}, E_{ele} and E_{vdW} are the bonded, electrostatic and van der Waals terms, respectively. G_{Pol} is the polar solvation free energy and $G_{np}$is the non-polar solvation free energy.

For each binding site identified during our long timescale and splitting protocols, we ran an ensemble of 25 4ns trajectories. Since the predominant ligand binding sites and poses were identified as the final frames from which subreplicas were initiated in the splitting protocol, we used these configurations as

the starting structure for ESMACS calculations performed for the ADRP system.

Our choice of running 25 simulations of 4 nanoseconds each is in accordance with previous findings by our group showing that 25 replicas of 4 ns are sufficient to obtain converged values of ΔG_{ESMACS}[22]. These trajectories were post-processed in MMPBSA.py to produce 25 binding free energy estimates, one for each replica within the ensemble. The reported ΔG_{ESMACS} is the mean of the sampling distribution of means for this sample of 25 free energy estimated obtained using bootstrapping. The associated error bars are the corresponding standard errors.

3.4.2 "Direct" binding free energy calculations

The "direct" binding free energy calculation method was originally developed by De Jong *et al.*[85] and later applied to 10 microsecond trajectories by Pan *et al.* in 2017[27]. We would like to point out that the method is justified on the basis that a sufficiently long single trajectory can be averaged to produce a meaningful macroscopic free energy. However, we will demonstrate that this assumption is not valid, and hence free energies computed through this method using a single trajectory are not reliable. To calculate the binding free energy, we use the following equations which were derived via statistical mechanics by De Jong *et al.*[85]

$$K_A = \frac{P_b}{P_u} v c^o N_{Av} \,;$$

$$\Delta G_b = -k_B T ln K_A \,.$$

Here, P_b and P_u are the fraction of simulation time in which the ligand is bound and unbound to the binding site of interest respectively, v is the volume of the simulation box (L), c^o is the standard-state concentration (1 mol L^{-1}), N_{Av} is Avogadro's number, k_B is Boltzmann's constant and T is the temperature (K). We define the ligand to be in the bound state when the first two closest distances between the heavy atoms of the ligand and the side chain heavy atoms of the binding site residues is <5 Å. All other frames are defined as unbound.

3.5 The Kolmogorov–Smirnov test

To compare the ligand-residue contact frequency distributions, we perform the pairwise Kolmogorov–Smirnov (KS) test. The test compares the underlying continuous distributions $F(x)$ and $G(x)$ of two independent samples (in this case, two ligand-residue contact frequency distributions, each derived from separate MD trajectories). Since the test is non-parametric, it is particularly suited to the comparison of ligand-residue contact frequency distributions as they have multiple peaks and are not normally distributed.

To test the statistical certainty of two distributions being different from one another, we use the two sided p-test. For this test, the null hypothesis is that both of the distributions are sampled from the same underlying distribution. All KS tests are computed using the SciPy package in Python[88].

4 Results and Discussion

This section is divided into two subsections. In the first subsection, we discuss aspects of our results that are important from the point of view of developing new scientific methods that yield statistically robust and reliable outcomes. We report our findings on the effect of stochasticity in MD simulations at "long" timescales. We show how this intrinsic characteristic of MD can be used to our advantage in order to enhance the sampling of phase space through introduction of biases. Further, we determine binding affinities using two different methods and compare them to discuss the advantages and disadvantages of each method and highlight scenarios where a particular method should be preferred. In the second subsection, we discuss the important scientific findings of our study. We describe the allosteric mechanism through which LZE binding affects the binding of RQN and show how the binding poses/sites are affected by this.

4.1 Development of Scientific Methods

4.1.1 Aleatoric uncertainty in "long" MD simulations

We have shown that classical molecular dynamics simulations are extremely sensitive to their initial conditions given their chaotic nature due to which two independent MD trajectories diverge exponentially with time[19]. This has been exhibited in several published studies for short simulations (up to a few nanoseconds) including ours[18,20,35]. Unprecedentedly, in this study we provide evidence for such divergence between independent simulations extending up to 10 microseconds. Our results conclusively show that MD trajectories lead to very distinct regions of a given phase space even when they are considered "long". Thus, results based on one-off "long" simulations are as unreliable as one-off "short" simulations. Indeed, it is essential to perform ensembles in all cases to quantify the uncertainty and ensure reproducibility of results. This is due to the mixing nature of the dynamics which is a necessary and sufficient condition to reach equilibrium[19,23].

Table 2 provides the number of binding sites sampled by the entire ensemble of 10 or 12 replicas for each system (ten for ADRP-tofacitinib complex and twelve for all other systems) in column 2. In the third column, it also includes the number of replicas that visit each binding site for each system. It is evident that not all sites are sampled in all simulations. There is substantial variation in the

Table 2: Sampling frequency of the different binding sites across all "long" independent replicas. The middle column shows the number of different binding sites sampled across all replicas for a given system. The last column shows an ordered set of the number of replicas visiting a given binding site for all sites in the middle column. The number in bold font corresponds to the experimental binding site. Note that this only captures whether a replica samples a given binding site at all or not. It does not take into account the amount of time spent at a given binding site by any replica. The total number of replicas is 10 for the ADRP system whereas 12 for all others.

System	# of binding sites	# of replicas visiting each site
ADRP-tofacitinib	4	3,4,**9**,6
PLPro-GRL	15	1,1,1,1,1,1,1,1,2,1,**1**,1,1,1,1
3CLPro-93J	13	1,**8**,1,2,1,1,1,3,3,1,1,1,1
3CLPro-RQN	9	7,2,1,**10**,2,1,1,1,1
3CLPro-RQN (with LZE)	12	6,4,2,2,**3**,**2**,4,3,1,2,1,1
3CLPro-LZE (with RQN)	8	6,**6**,**5**,3,4,3,4,1

binding sites sampled both across replicas for each system as well as across all systems studied. For instance, in the case of the ADRP-tofacitinib system, four different binding sites are sampled by 3, 4, 9 and 6 replicas respectively. Comparing this behaviour with that of the other four systems studied (all relatively bigger in size), we can clearly see that they differ in that the number of sites observed is much higher with the number of replicas visiting each site being smaller. Taking the example of the PLPro-GRL system, there are 15 different binding sites observed with each only sampled by a single replica for all but one site. Furthermore, 9 out of 12 replicas exclusively sample only a single binding site. This behaviour is in contradiction to what we see for the ADRP-tofacitinib system exhibiting the extent of variation in sampling that may be observed across different systems using an ensemble of long independent simulations. The sampling behaviours of the other three systems fall in between the two extremes discussed above. It should be noted here that, in the above analyses, a replica is considered to have sampled a given binding site only if its ligand fractional occupancy is ≥ 0.03 around that site. In other words, a replica is assumed to have sampled only those binding sites that appear in the volume occupancy maps (and have non-negligible peaks in the corresponding contact frequency distributions) displayed in Figures 3 and S1-S4. It is possible that a replica has visited other binding sites too for a very short period of time but such transient events are neglected in our analyses as such a binding process cannot be considered stable.

There is also a non-negligible variation in sampling across replicas for each system studied. Taking the ADRP-tofacitinib system as an example, as already noted, four different binding sites (denoted as A, B, C and D) have been sampled collectively by the ensemble of ten 10 μs long replicas. The crystallographic site (C) is sampled by 9 out of 10 replicas whereas all other sites are only located by a

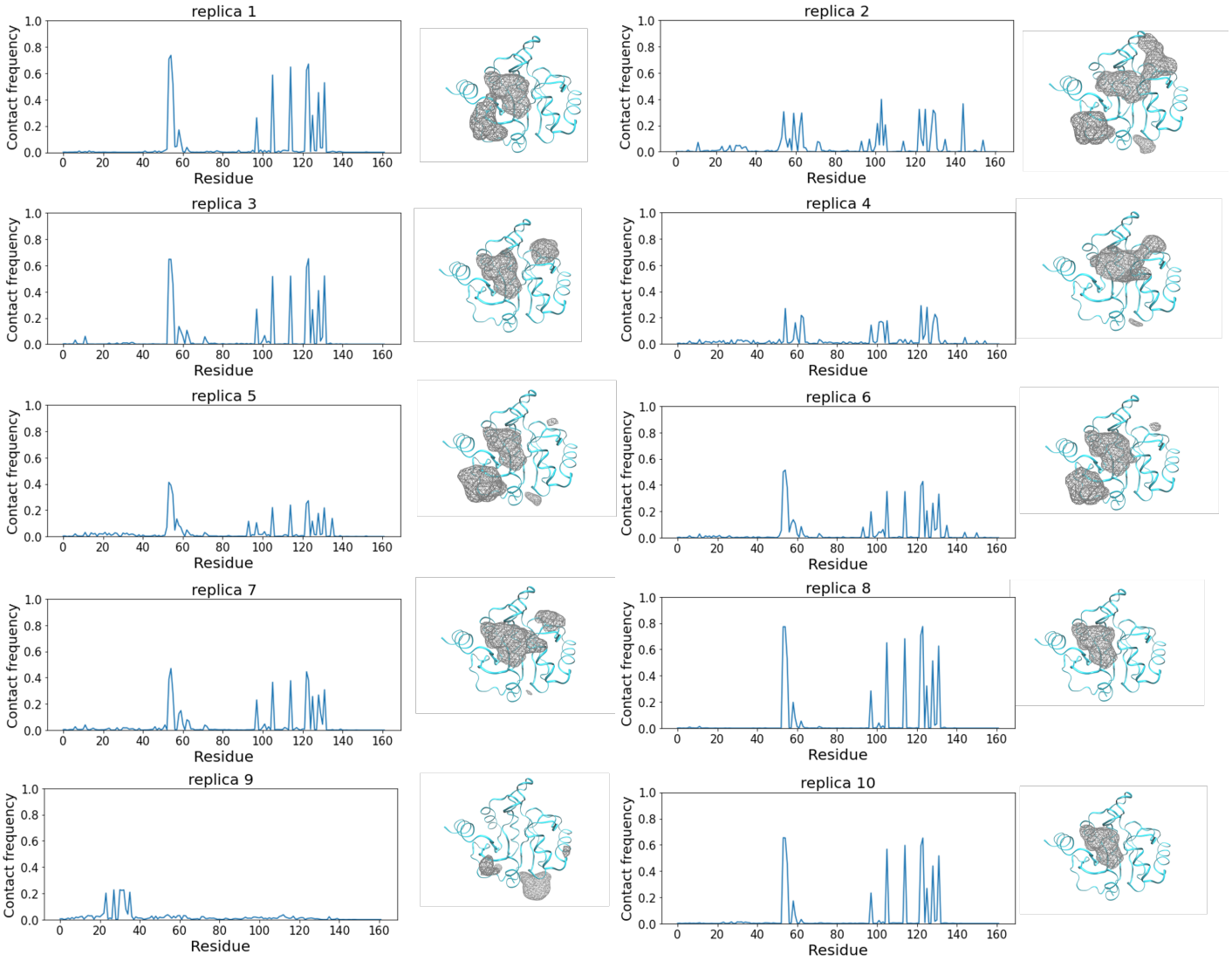

Figure 3: Tofacitinib (ligand) - ADRP (protein) residue contact frequency distribution plots for each "long" replica are shown adjacent to their respective ligand occupancy maps. The ligand-residue contact frequencies correspond to the fraction of frames in which a hydrophobic contact is formed between the ligand and a given protein residue. Occupancy maps of tofacitinib around the ADRP protein represent the isovalue surfaces (wireframe representation) rendered at the fractional occupancy of 0.03 across all frames of the simulation trajectory. In other words, they represent volumes of the simulation box where the ligand is likely to be found with 97% probability, that is in 97% of all trajectory frames.

Table 3: Mean and range of KS statistics values across all replicas for all systems studied. The number of KS values ≥ 0.2 (an arbitrary threshold) as well as the number of p-values ≥ 0.05 (in percent terms).

System	Mean	Range	KS ≥ 0.2	p-value ≥ 0.05
ADRP-tofacitinib	0.26	0.11-0.54	82.2	6.7
PLPro-GRL	0.32	0.13-0.59	95.4	0
3CLPro-93J	0.31	0.15-0.61	97.0	0
3CLPro-RQN	0.26	0.11-0.52	78.8	1.5
3CLPro-RQN (with LZE)	0.33	0.10-0.66	86.4	1.5
3CLPro-LZE (with RQN)	0.23	0.12-0.50	75.8	0

smaller number of replicas. There are two replicas (IDs 8 and 10) that exclusively sample site C, whereas two other replicas (IDs 2 and 5) sample all four sites. The remaining six replicas sample different subsets of the four binding sites in different combinations and proportions. It should also be noted that the sampling of the ligand around site C is quite different across each of the 9 replicas as quantified in the following paragraphs. Similar behaviour applies to all other systems studied.

In order to provide a visual representation of the sampling variation discussed above, we have calculated the contact frequency distributions for individual replicas (refer to section 3 for details). Figure 3 displays contact frequency distributions of the ADRP-tofacitinib complex for each of the ten long (10 μs duration) simulations along with corresponding volume occupancy maps. It can be clearly seen that the signature frequencies of site C are visible in the contact frequency plots of all but one replica (only replica ID 9 does not sample the crystallographic site). Replica IDs 8 and 10 exclusively sample site C and hence display identical peak distributions whereas other replicas have different peak distributions due to overlapping frequencies from other binding sites samples. Similarly, replica ID 9 predominately samples site A, clearly showing the corresponding signature frequencies. Another replica that has a non-negligible peak at site A frequencies is replica ID 2 as is also confirmed by the corresponding volume occupancy maps. It should be noted that the magnitude (peak heights) of these signature frequencies for different binding sites are different across replicas. Similar figures with contact frequency distributions of the other systems studied have been included in the Supporting Information (Figures S1-S4) which all convey the same message as above.

Quantification of Aleatoric Variability

In order to derive robust insights, it is essential that we quantify the extent of variability between the long replicas so as to determine their reproducibility. To achieve this, we compute pairwise Kolmogorov-Smirnov (KS) test statistics for each pair of the long replicas. The pairwise KS statistic has a range

of 0 to 1, where 0 indicates that the two sample distributions being compared are sampled from an identical underlying distribution, and 1 indicates the converse case. Figure 4A exhibits a matrix of pairwise KS statistics for 45 possible pairs of replica trajectories for the ADRP-Tofacitinib complex. The resultant values fall between a wide range of 0.11 to 0.54. However, 40 of them are ≥ 0.15 and 37 are ≥ 0.2. The mean value of the KS statistic for all 45 pairs is 0.26. Figure 4B displays a matrix of corresponding p-values from pairwise KS statistics. A p-value of < 0.05 signifies that the null hypothesis (that the two underlying distributions are identical) can be rejected with 95% confidence. This, in turn, means that there is a 95% chance that the two samples compared are drawn from different underlying distributions. We obtain a p-value of ≥ 0.05 for only 3 out of the 45 pairs of replicas ($\sim 6.67\%$). Thus, 42 pairs ($\sim 93.33\%$) indeed sample very different regions of phase space. Figures similar to 4 for all other systems have been included in the Supporting Information (Figures S5-S9). In addition, Table 3 contains relevant statistics (as discussed above for the ADRP system) for all systems. All these figures and data show that the variability across replicas is prominent in all systems studied without exception even at the microsecond timescale.

Figure 5 displays the cumulative density functions (CDFs) of ligand-residue contacts for all ten replicas of the ADRP system. Constructs known as p-boxes (regions between two extreme CDFs) are often used to visualise how the distribution of outcomes is controlled by aleatoric and epistemic uncertainty[89]. It is clear from Figure 5 that the p-box generated by ten "long" independent replicas has a wide range - another representation of the extensive variation of sampling across replicas. Figures displaying the CDFs of ligand-residue contacts and corresponding p-boxes for all other systems have been included in the Supporting Information (Figures S10-S14) with identical observations.

The above findings establish beyond doubt the non-reproducibility of long trajectories. They confirm that it is far-fetched to draw final conclusions on the true nature of a system from an individual MD simulation, regardless of its temporal duration[19]. Indeed, this is a direct reflection of the chaotic nature of molecular dynamics simulation and, from a theoretical standpoint, shows that individual 10 microsecond trajectories can never be used to determine equilibrium behaviour. In fact, equilibrium is meaningful only for ensembles of trajectories which manifest the required dynamical instability. While individual trajectories are time reversible, the approach to equilibrium is a probabilistic property of ensembles which requires the dynamics to be chaotic[19].

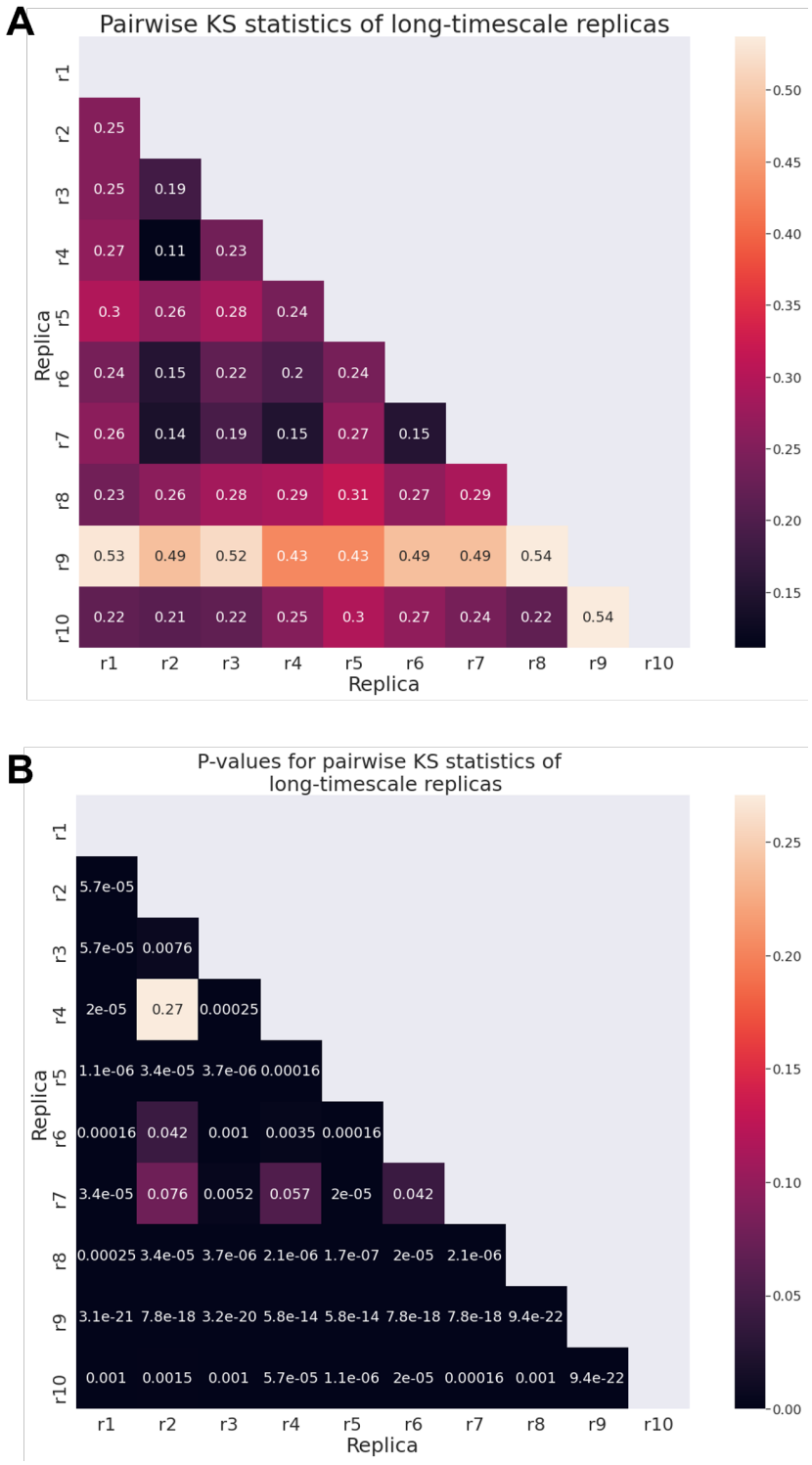

Figure 4: Heatmaps depicting pairwise Kolmogorov-Smirnov (KS) statistics for all pairs of 10 microsecond replicas for ADRP-tofacitinib complex: (A) Pairwise KS-statistics for each pair of long replicas, (B) Two-tailed p-values corresponding to each KS-statistic result. Values of < 0.05 allow the null hypothesis (that the underlying distributions are identical) to be rejected with 95% confidence.

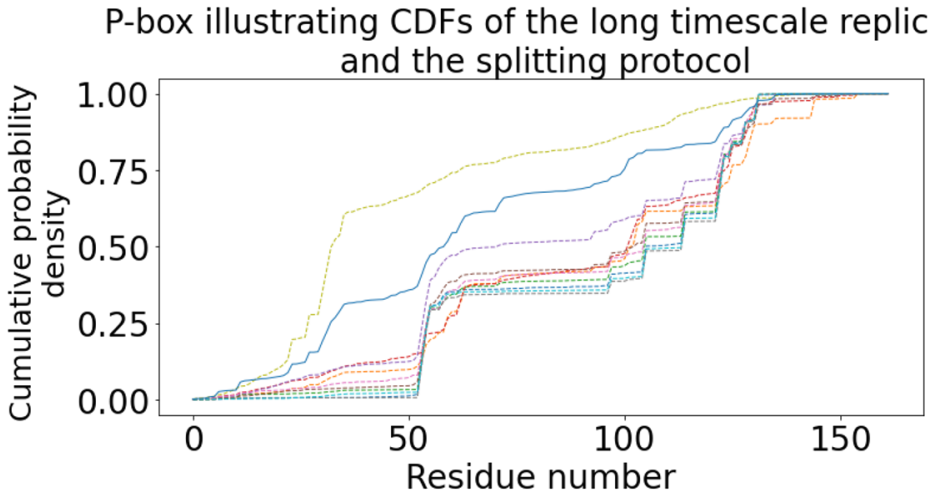

Figure 5: Cumulative density functions (CDFs) of the contact frequency distributions for all "long" replicas (dashed lines) as well as concatenated splitting protocol trajectories (solid line) for ADRP-tofacitinib system. The width of the p-box so generated indicates the extent of variability across "long" replicas compared against the splitting protocol.

Variability in Free Energy Estimates

Free energy is a thermodynamic observable of importance for protein-ligand complexes in the drug discovery context. Therefore, we also look at the extent of variation in free energy estimates obtained using independent "long" replicas of MD simulations. We used ΔG_{direct} as a measure of absolute binding free energy which was originally developed by De Jong *et al.*[85], and later applied to 10 microsecond trajectories by Pan *et al.* in 2017[27] (details in section 3). In Pan *et al.*'s study, the authors reported "converged" ΔG_{direct} values from individual trajectories, where binding free energies were considered converged if the difference between the estimated binding affinity and its final value as a function of simulation time was in the range of $\pm$ 0.5 kcal/mol. Furthermore, the error associated to these binding free energies was computed by calculating the variation of ΔG over blocks of 2 μs of simulation time. In the present work, we demonstrate that ΔG_{direct} varies substantially between separate independent long timescale replicas, and hence once again individual "long" simulations do not provide reliable binding free energy estimates. The salient point here is that, contrary to received wisdom in the literature on molecular dynamics, averaging over an individual long timescale simulation is not equivalent to averaging over an ensemble of simulations. Indeed, thermodynamic quantities arise from ensemble averaging in statistical mechanics and unless one averages over a timescale of the order of a Poincaré recurrence, a one-off MD trajectory will produce the wrong results[19]. Compounding this, a one-off simulation does not provide the means to compute precise results or conduct meaningful uncertainty quantification.

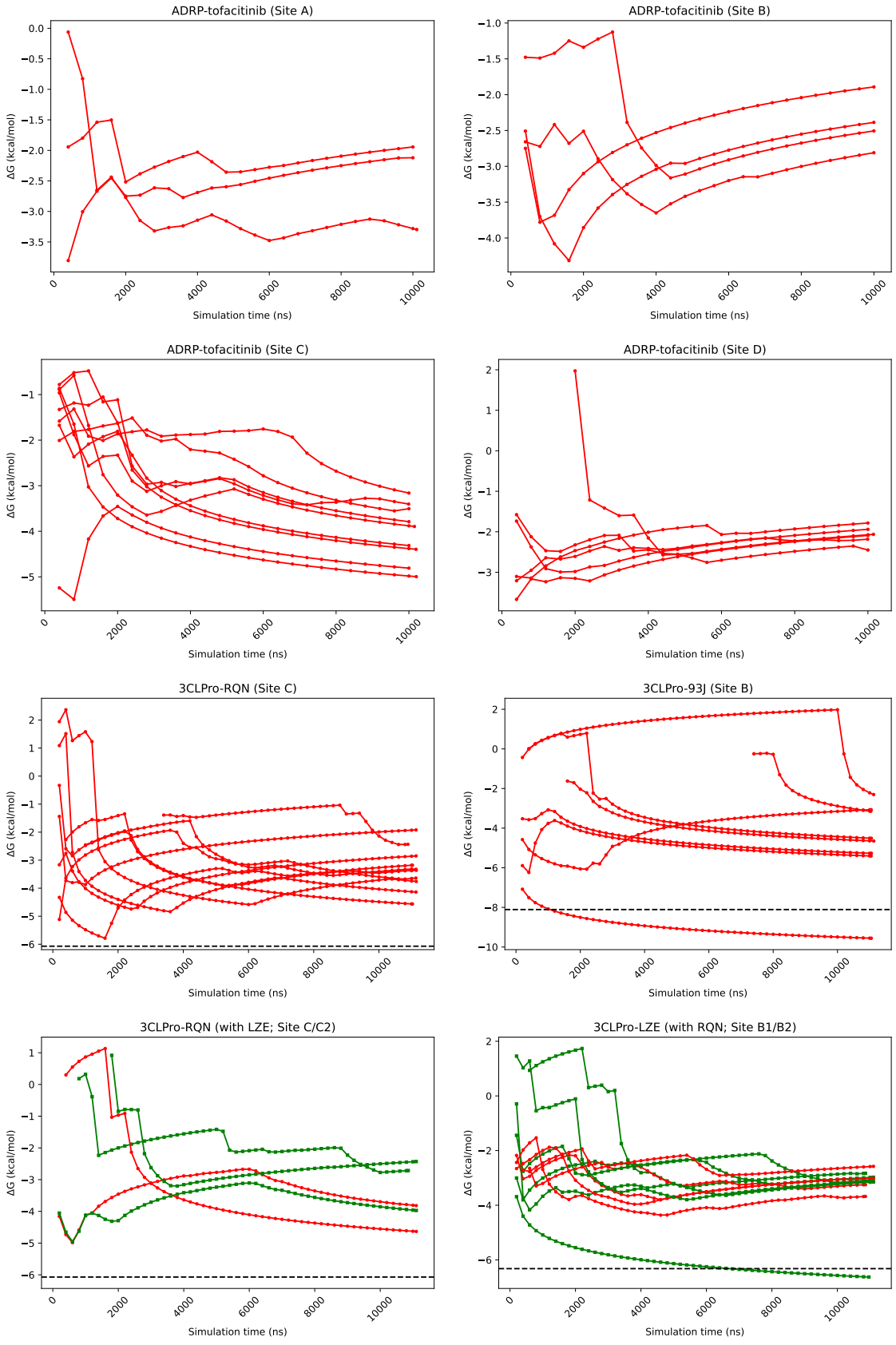

Figure 6: Running averages of ΔG_{direct} for tofacitinib binding to ADRP at all of the identified ADRP binding sites (top two panels) and for the experimental binding sites of 3CLPro systems (bottom two panels) for all ten or twelve 10μs trajectories. The horizontal black dashed lines in the plots of 3CLPro systems correspond to the respective experimental binding affinities.

Table 4: Mean and spread (that is difference between extreme values) of ΔG_{direct} across all replicas for the binding site that is visited by the most number of replicas for each system studied. Note that such a binding site is not always the experimentally determined one. Error bars are the standard errors. All values are in kcal/mol.

System	Mean	Spread (# of replicas)
ADRP-tofacitinib	-4.03(0.20)	1.83(9)
PLPro-GRL	-3.47(0.27)	0.77 (2)
3CLPro-93J	-4.74(0.74)	7.26 (8)
3CLPro-RQN	-3.32(0.23)	2.64 (10)
3CLPro-RQN (with LZE)	-3.95(0.32)	2.16 (6)
3CLPro-LZE (with RQN)	-3.63(0.55)	3.80 (6)

Figure 6 shows the running averages of ΔG_{direct} for all four binding sites of the ADRP system (top two rows) as well as for crystallographic sites of all 3CLPro systems (bottom two rows) from all replicas that sample them. The inter-replica variation is clearly visible from these plots for all systems. This variability shows that results obtained from individual "long" trajectories are not reproducible or precise. Furthermore, it is not able to reliably predict binding free energies with chemical accuracy ($\pm$ 1 kcal/mol). Figures displaying running averages of ΔG_{direct} for all binding sites of all systems studied have been included in the Supporting Information (Figures S15-S19). All of them show behaviour similar to that discussed above in terms of ΔG variability. Table 4 includes the mean ΔG_{direct} values along with error bars for the most frequently visited binding site (which is not the experimental binding site for PLPro-GRL and 3CLPro-RQN (with LZE)) across all replicas for each system. It also provides the spread (that is, the difference between the two extreme values) for each such binding site which is around 2-3 kcal/mol for most cases but can be as high as 7 kcal/mol (for instance 3CLPro-93J). Another point worth noting from Figures 6 and S15-S19 is that the first binding event occurs at varying time durations across ensemble members as captured by the different onset simulation times of the running average plots. This confirms that the dynamical behaviour has substantial variability at the microsecond timescale for molecular dynamics, just as it does for shorter time scales.

To obtain meaningful estimates of ΔG_{direct} we must take into account the results from all members of an ensemble. To do this, we employ bootstrapping to obtain sampling distributions of the mean for ΔG_{direct} by resampling 5000 times with replacement. The original sample used for such analysis is the ensemble of ΔG_{direct} values from all replicas that sample a given binding site. The probability density functions of the sampling distributions of means so obtained are displayed in Figure 7 for a selection of systems studied. The corresponding underlying distributions, that is the frequency distributions of the original sample of ΔG_{direct} values used to perform bootstrapping, are shown in Figure 8. As we have

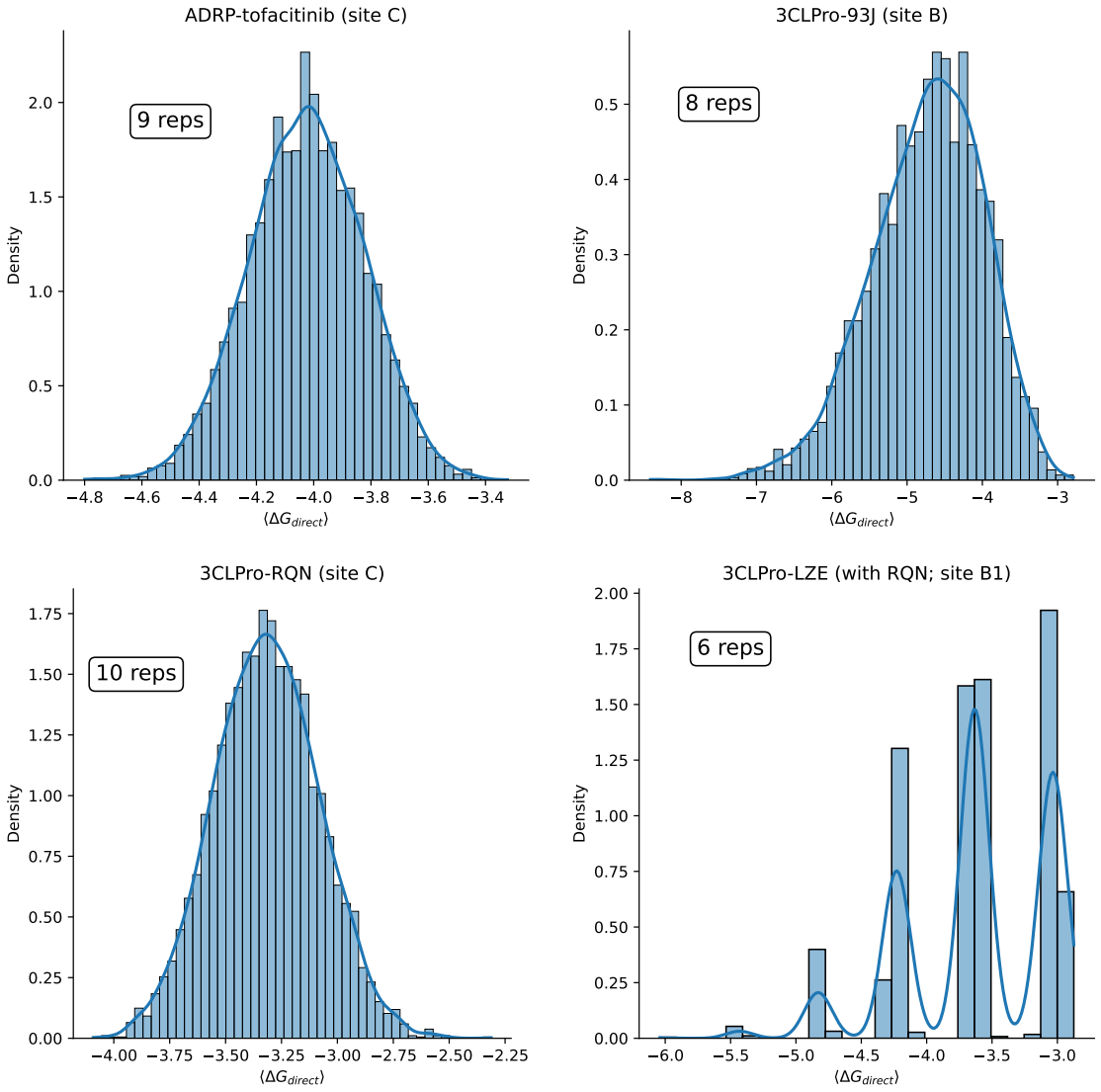

Figure 7: Probability density functions of sampling distributions of mean direct free energies ($\langle \Delta G_{direct} \rangle$) obtained with bootstrapping (5000 resamples) for four of the systems studied at their respective crystallographic binding sites. Bar plots display density histograms whereas solid lines represent the respective kernel density estimations. The original sample used for bootstrapping in each case is the set of final ΔG_{direct} values from all replicas that sample the respective binding sites. Sample sizes are shown in the text boxes within each plot. Given the small sample size, not all distributions shown here are Gaussian. The x-axis is expressed in kcal/mol.

shown in previous studies for relatively short duration MD trajectories, it is possible that the underlying free energy distributions may be non-Gaussian whereas the corresponding bootstrapped distributions approach the Gaussian functional form with increasing sample size as a consequence of the central limit theorem[18,23,24,35,90–93]. Figures 8 and 7 provide evidence of similar behaviour in case of "long" MD simulations as well, although given the small sample sizes (as shown in inset) not all bootstrapped distributions are Gaussian either. To be sure, an ensemble of size ≤ 10 is far too small to draw definitive conclusions on the true form of the underlying distribution. To ensure convergence of ΔG_{direct}, it would be necessary to determine the change in the bootstrapped value of ΔG_{direct} as a function of the number of replicas. Upon convergence, the estimate for the binding free energy could be classified as reliable and reproducible.

The crucial idea here is that to even begin to generate reproducible estimates for ΔG_{direct}, running ensembles of simulations, irrespective of their length, is an imperative. Interestingly, the non-normal nature of free energy distributions implies more frequent occurrence of outliers than would be expected with normal distributions that necessitates relatively more data in order to obtain reliable estimates. Since the variability that exists across replicas within an ensemble of simulations is caused by the intrinsically chaotic nature of MD simulations, these principles will apply to the calculation of any MD derived macroscopic expectation value.

4.1.2 Biased versus Unbiased Sampling

In the previous section, we have described results from unbiased MD simulations and shown that the sampling may vary substantially on repeating a simulation. In this section, we include results from the biased simulation protocol named the "splitting protocol" which involves biasing the sampling of phase space towards sites of interest (described in detail in section 3). It should be noted that, in principle, such splitting steps can be continued further until the desired level of sampling has been achieved (for instance, if there is a substantial variation in the binding poses across sub-replicas that need to be explored further, and so on).

In the case of ADRP-tofacitinib, a highly multi-modal distribution is observed across the initial 20 200ns replicas with multiple binding sites explored as shown in Figure 9. It can be seen that all four binding sites are identified while each replica possesses a unique distribution of tofacitanib-residue contact frequencies. Identical figures for all other systems studied have been included in the Supporting Information (Figures S20-S24) with similar conclusions. For the ADRP system, the subreplicas were initiated from the final frame of replicas 1, 9, 14, 15 and 20, where the ligand was positioned at binding

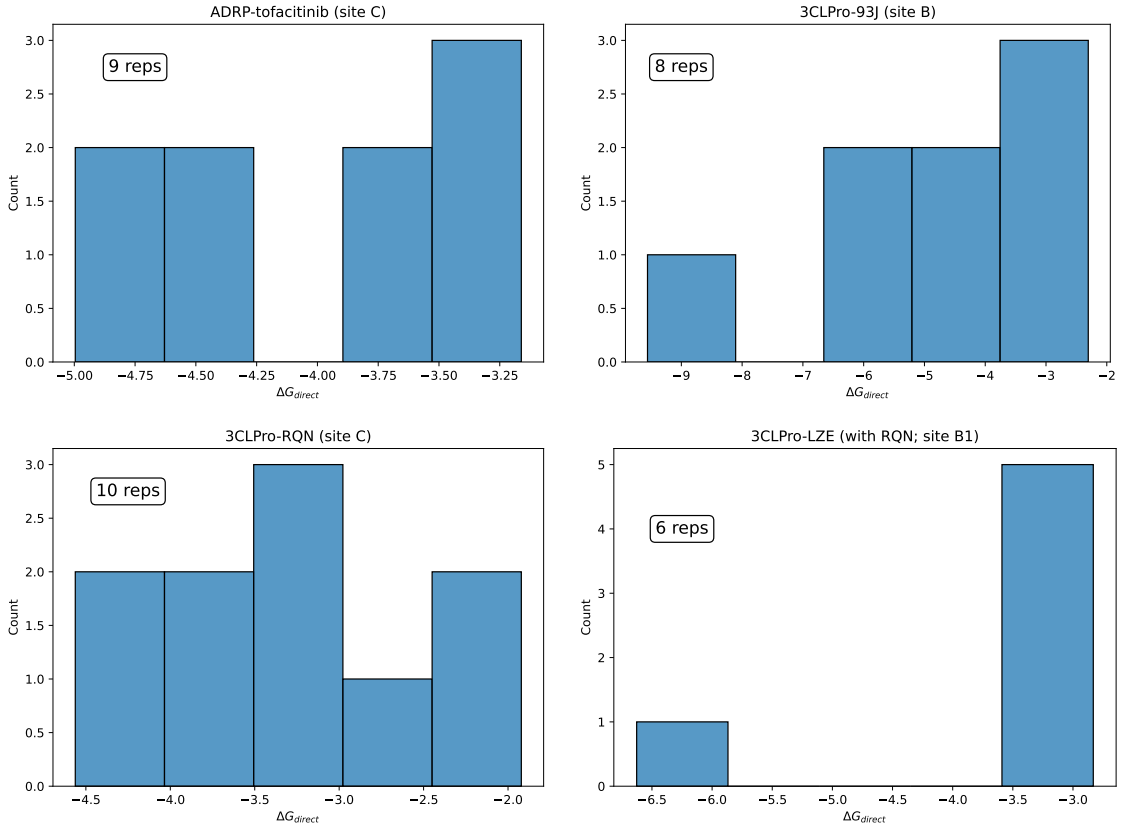

Figure 8: Frequency distributions of ΔG_{direct} values (using 5 bins) for four of the systems studied at their respective crystallographic binding sites obtained from independent "long" MD trajectories. The number of replicas sampling the binding site in each case is shown in the text box on the respective plots. The data suggest that there may be non-Gaussian behaviour in the underlying distribution. The x-axis is expressed in kcal/mol.

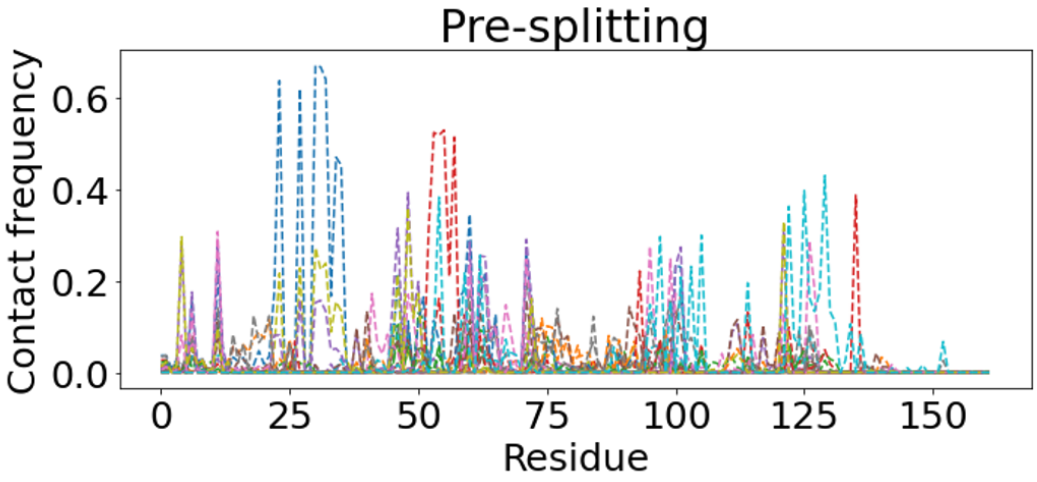

Figure 9: Ligand-residue contact frequencies for the initial twenty 200 nanosecond trajectories under splittling protocol for ADRP-Tofacitinib complex.

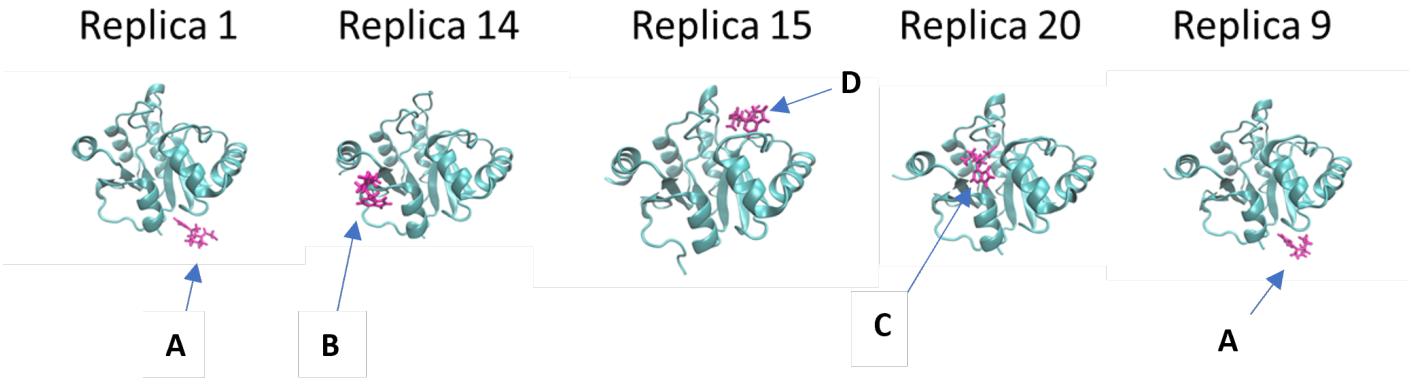

Figure 10: Structures of ADRP-tofacitinib complexes from which sub-replicas are initiated. Each annotation corresponds to a binding site identified by the splitting protocol.

sites A, B, C and D as shown in Figure 10.

As noted earlier, site C is the crystallographically defined site of ADP ribose and has the most negative binding affinity for tofacitinib. This relatively high thermodynamic stability at site C is also reflected in the occupancy maps and ligand-residue contact frequency distribution plots of sub-replicas initiated from the end frame of replica 20 as shown in Figure 11. The ligand possesses a well-defined pose across all replicas when initiated from site C (sub-replicas of replica 20). This is contrary to the behaviour seen when initiating sub-replicas from the ligand located in other binding sites where the ligand explores multiple sites over each set of sub-replicas as also evident from Figure 11. Overall, this provides evidence that tofacitinib would act to competitively inhibit the protein.

The important thing to note here is that the aleatoric nature of MD has been utilised to our advantage to substantially accelerate the exploration of phase space by introducing appropriate bias to the sampling

Table 5: Mean and range of KS statistics values comparing splitting protocol with each long replica for all systems studied. The number of KS values ≥ 0.2 (an arbitrary threshold) is given in percent terms.

System	Mean	Range	KS ≥ 0.2
ADRP-tofacitinib	0.31	0.25-0.36	100
PLPro-GRL	0.30	0.21-0.47	100
3CLPro-93J	0.32	0.21-0.48	100
3CLPro-RQN	0.21	0.14-0.39	58.3
3CLPro-RQN (with LZE)	0.42	0.13-0.6	83.3
3CLPro-LZE (with RQN)	0.26	0.18-0.44	83.3

compared to a single simulation of the duration given by the aggregate time of all runs under the splitting protocol in a much shorter wall-clock time. In the following paragraphs, we further substantiate this point by comparing the results from biased and unbiased sampling.

First of all, we compare the contact frequency distribution of the aggregated (biased) sampling using the splitting protocol (9 μs) with those from all the individual unbiased sampling from the 10 long simulations (10 μs each). Figure 12 displays such comparisons for ADRP-tofacitinib complex (refer to Figures S25-S29 in the Supporting Information for all other systems). It is evident that each 10 microsecond replica samples a mere subset of possible binding sites explored across the 9 microseconds of the splitting protocol. The only exception to this general observation are the contacts of the ligand with residues 140-160 that are exclusively observed in long simulations. We will discuss this exception below. The significant difference in sampling between the splitting protocol and each individual long replica is quantified with two-sample KS statistics as shown in Figure 13 for ADRP system (and Figure S30-S34 for all other systems). For ADRP system, KS statistics varies between 0.25 and 0.36 with an average of 0.31 and the corresponding averaged p-value is 1.37×10^{-5}. Table 5 shows mean and range of KS statistics values for other systems. They fall in a similar range going as high as 0.6 and as low as 0.13 in some cases.

Figure 5 displays a p-box for the contact frequency distributions for the splitting protocol as well as all long replicas separately for ADRP system (and Figures S10-S14 for other systems). It clearly shows the bounds on the cumulative probability of the ligand contacting a given residue across the full set of simulations furnishes a clear visualisation of the aleatoric uncertainty that is associated with the ligand-residue contact frequency across all simulations.

On carefully observing Figure 12, it can be noted that when considering all long trajectories in aggregate, we are able to recover all ligand-protein interactions which are identified across the splitting protocol. To see this more clearly, we plotted the ligand-residue contact frequency distributions for

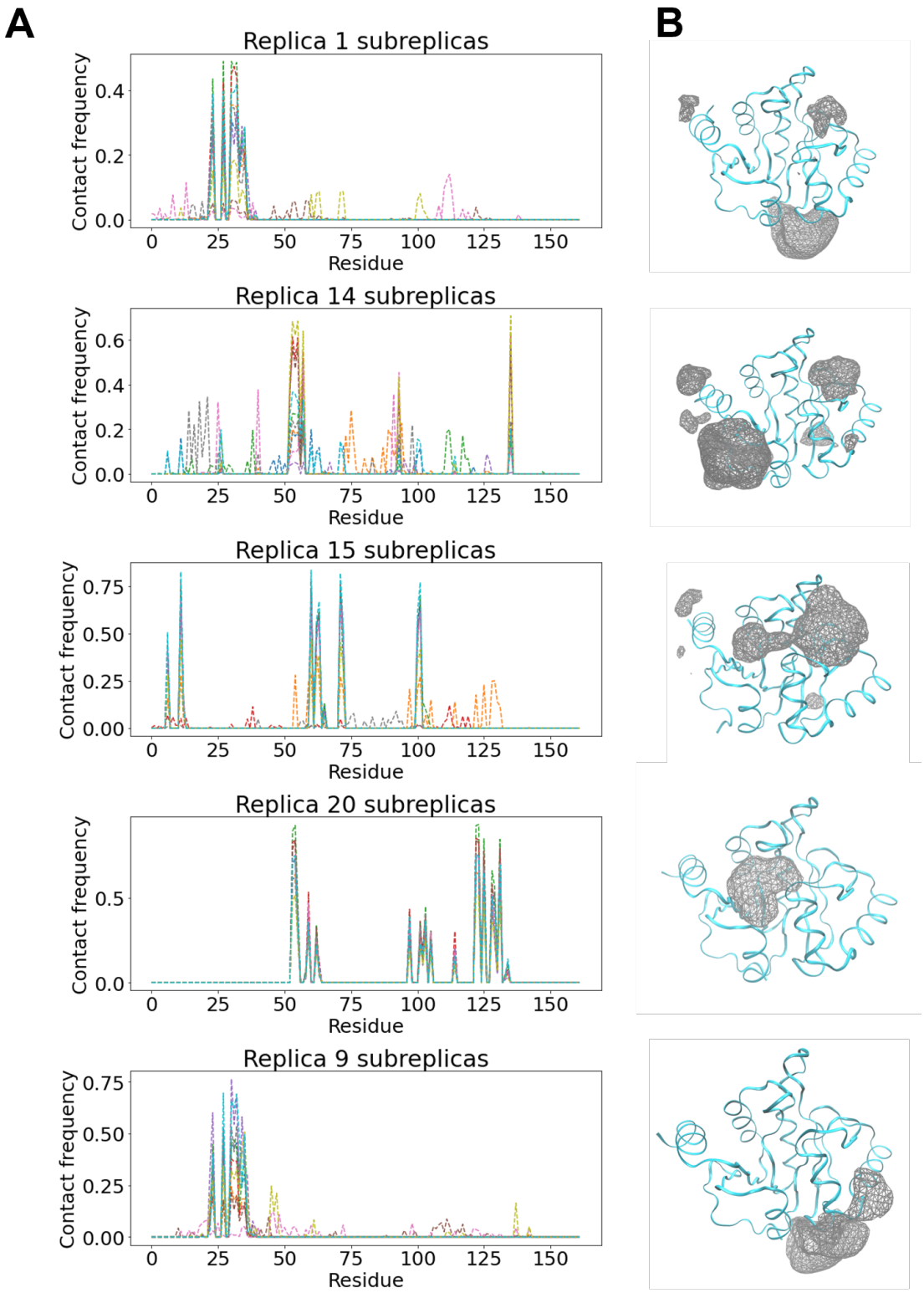

Figure 11: ADRP-tofacitinib complex with splitting protocol: (A) Distribution of ligand-residue contact frequencies for each set of sub-replicas, (B) Volume occupancy maps of the ligand around protein rendered at an isovalue of fractional occupancy 0.03. For each set of subreplicas, the wireframe isosurface represents the area of the simulation box where the ligand is likely to be found with 97% probability.

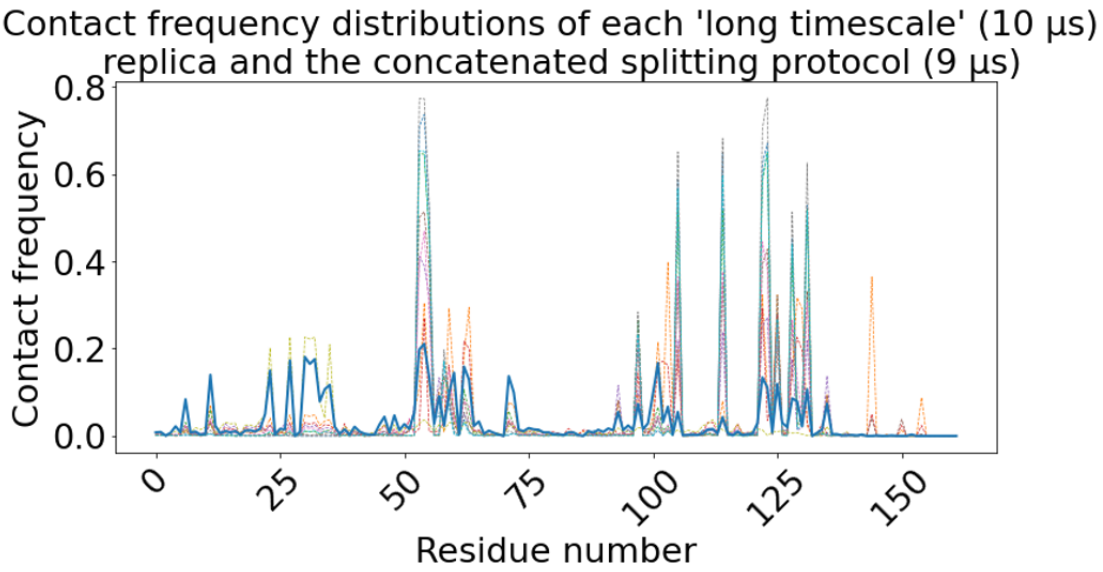

Figure 12: Contact frequency distributions of all "long" (10 μs) replicas (dashed lines) compared to that of the splitting protocol (9 μs) (solid blue line) for ADRP-tofacitinib complex.

aggregated sampling time from both the splitting protocol (9 μs in total) as well as unbiased sampling (100 μs in total) for ADRP system in Figure 14. Here we see that the concatenated trajectory reproduces all modalities which occur across the splitting protocol, albeit with different statistical weights. Incidentally, if our aim is to simply explore all possible binding sites and dominant poses within those sites, the unbiased long timescale protocol is far less efficient than the splitting protocol which achieves this aim in an aggregate of 1 day and 7 hours of wall clock time rather than 43 days 2 hours of wall clock time required with the former. However, care must be taken when thermodynamic quantities need to be evaluated/predicted using the splitting protocol as the biased sampling leads to biased weights of the microstates sampled that may affect the averages obtained.

Nevertheless, there are advantages to performing an ensemble of long simulations rather than the splitting protocol. Namely, there are key poses and contacts identified during the long timescale protocol which are highly unlikely to be observed by the shorter timescale splitting protocol. In Figure 15, we see three tofacitinib-ADRP contacts which were observed during the long timescale protocol, but not during the splitting protocol. These contacts occur with residues ASN37, LEU53 and VAL36. Upon inspection, we find that all three residues are buried deep within site C (the crystallographically determined binding site). This indicates that a long duration of wall time is typically required in order to explore these "rare" poses as access is required to more buried regions of the ADRP active site.

Until now, we have discussed the variation across long timescale MD trajectories and emphasised that ensemble simulations are necessary for UQ irrespective of the duration of simulation. However, as

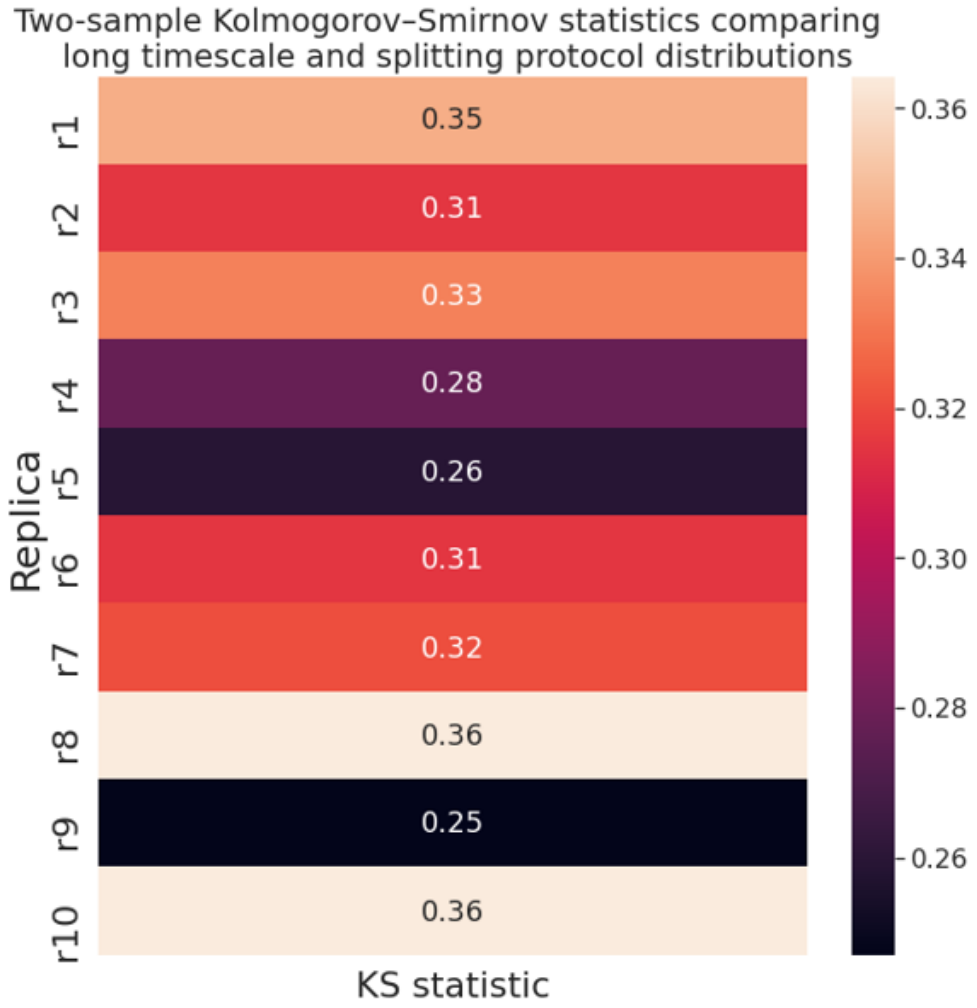

Figure 13: Two-sample KS statistics comparing the contact frequency distributions of the concatenated splitting protocol to those of each "long" replica for ADRP-tofacitinib complex.

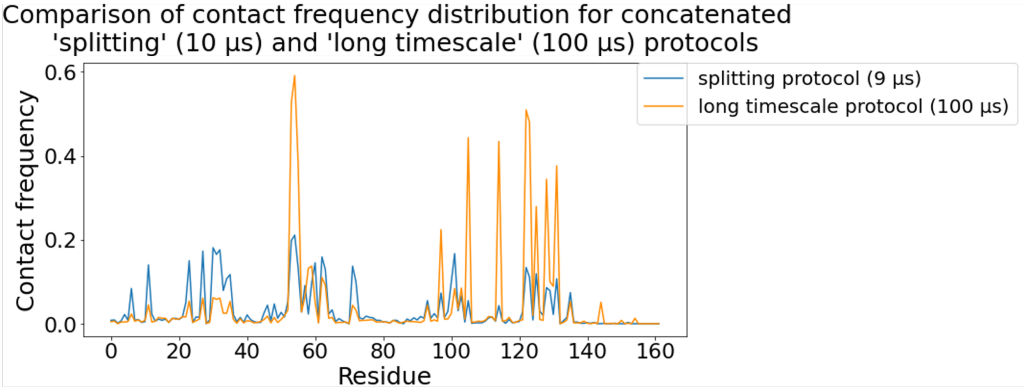

Figure 14: Contact frequency distribution of the concatenated "long-timescale" protocol (100 μs) compared to the contact frequency distribution of the splitting protocol (9 μs) for the ADRP-tofacitinib system.

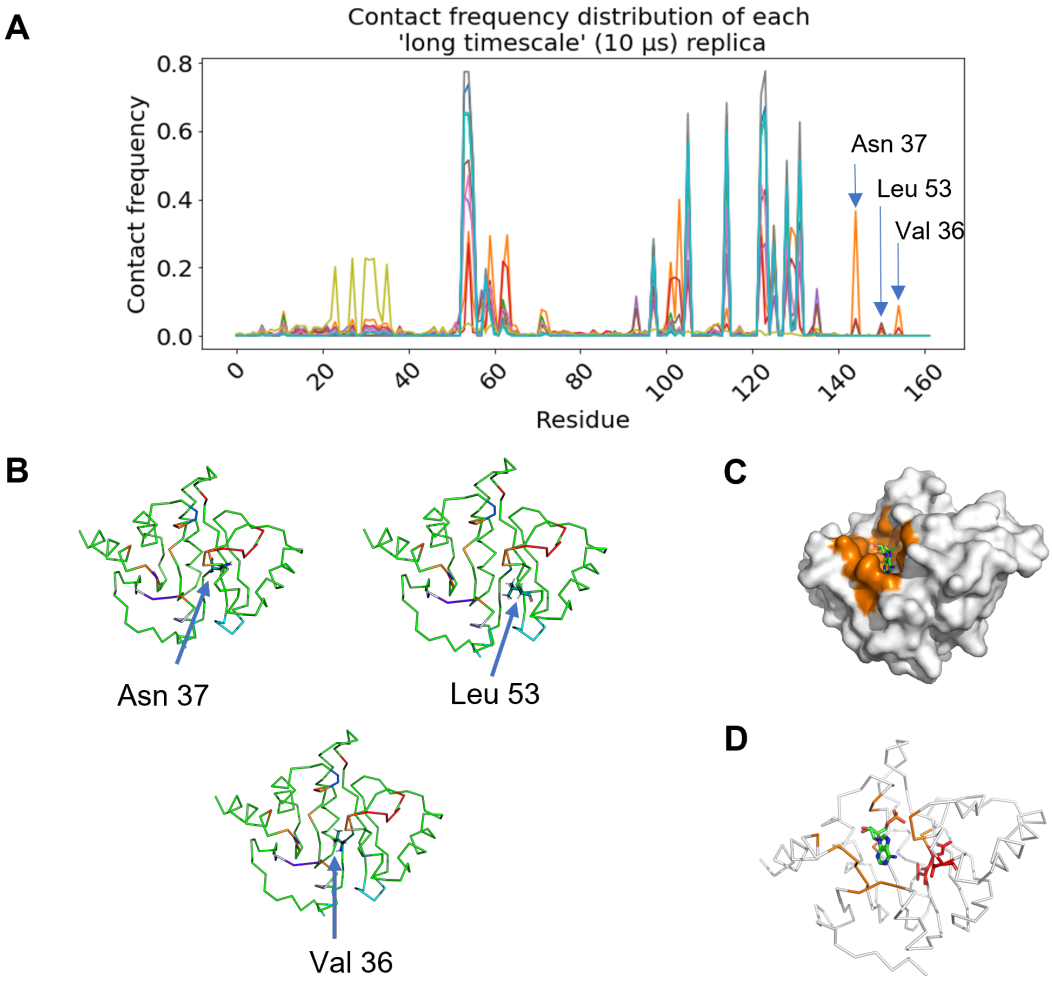

Figure 15: Tofacitinib-residue contacts identified only in the long-timescale protocol: (A) Tofacitinib-Residue contact frequency distribution plots of all ten 10 microsecond replicas, overlaid onto a single graph. Annotations clarify peaks which are unique to the long timescale protocol, namely, ASN 37, LEU 53 and VAL 36; (B) Ribbon representations of ADRP structures, with the location of each residue of interest annotated. Residues of interest have a 'liquorice' representation; (C) Surface representation of ADRP with the crystallographically defined pose of AMP in the binding site. The region defined by high-frequency residues is highlighted in orange; (D) Ribbon representation of ADRP with the crystallographically defined pose of AMP in the binding site. The region defined by high-frequency residues is highlighted in orange. The residues unique to the long-timescale protocol (ASN 37, LEU 53 and VAL 36) are highlighted in red.

already discussed in section 1, several accelerated sampling protocols (including the splitting protocol employed in this study) that are based on performing "ensembles" are also expected to exhibit similar variation and would require performing ensembles for UQ. We have already shown this for replica exchange methods in some of our previous works[21,66]. Nevertheless, this aspect has not been addressed adequately in the literature for other accelerated sampling protocols as the reported errors for such methods are all derived from the data generated from a single execution of the protocol, but never from ensembles comprising multiple instances. One reason of this shortcoming might be the computational cost associated with all these methods. We hope to return with a subsequent study where we will discuss this issue systematically.

4.1.3 Free Energy Methods: Direct versus ESMACS

In this section, we have compared free energies obtained from different free energy protocols. We have already seen ΔG_{direct} results for the different systems in previous sections. Now, we directly compare them to ΔG_{ESMACS} results obtained through the ESMACS protocol for ADRP-tofacitinib system. The standard ESMACS protocol (denoted as "ESMACS-s" and involves performing an ensemble typically of 25 MD simulations of 4 ns duration starting from a chosen conformation) has been extensively applied to a diverse range of protein-ligand systems and shown to rank ligands with very high precision[22,74,75]. In this study, we chose the most stable binding pose (the one with the least RMSD) from the different sub-replicas of the splitting protocol at each binding site as the starting structure for our standard ESMACS calculations. Table 6 and Figure 16 shows a comparison of ΔG_{direct} and $\Delta G_{ESMACS-s}$. We find that both methods achieve strongly correlated results with a correlation coefficient of 0.87 which indicates a very similar estimate of relative binding affinity for the ligand at each of the binding sites. However, it should be noted that ESMACS is not an accurate method and hence the absolute ΔG values from the two methods cannot be compared directly. These results clearly indicate that tofacitinib acts as a competitive inhibitor for ADRP by binding to the crystallographically resolved site.

The direct method involves a much larger amount of sampling as compared to ESMACS-s that involves performing short MD simulations of only a few nanoseconds duration. It is, however, notable that ESMACS-s is still able to obtain almost identical ranking of ligand-protein complexes with such little sampling which makes it a much more efficient method when accuracy is not necessary. Nevertheless, it is well known that ESMACS-s results depend heavily on the initial binding pose/structure of the ligand-protein complex being studied due to the short duration of simulations; this can be a drawback in some cases where the initial structure is not known correctly. For instance, when performing "long"

Table 6: Free energies obtained using different protocols. "ESMACS-s" and "ESMACS-l" correspond to free energies obtained using the standard ESMACS protocol (initiated from a chosen conformation) and those using bound conformations extracted from "long" trajectories, respectively. Error bars are included in brackets and denote the standard errors across all replicas that sample a given binding site. All values are in kcal/mol.

Binding site	Direct	ESMACS-s	ESMACS-l
A	-2.45(0.35)	-12.94(0.17)	-16.21(0.47)
B	-2.40(0.17)	-17.11(0.10)	-20.34(0.45)
C	-4.03(0.20)	-27.32(0.16)	-36.97(1.02)
D	-2.08(0.08)	-17.97(0.43)	-19.18(0.71)

simulations, a large number of binding poses are observed and it is hard to identify the most stable one in the absence of available experimental information. In such cases, ESMACS-s is not so useful as resultant ΔG values may vary substantially. As an example, we randomly picked out two different binding poses of 93J sampled within the long trajectories of 3CLPro-93J system at three binding sites (A, B and H1) and performed ESMACS-s calculations using each of them as the starting structures. The differences in $\Delta G_{ESMACS-s}$ values obtained starting from the two different binding poses at each 93J binding site are 11.44 kcal/mol, 4.93 kcal/mol and 6.61 kcal/mol respectively which are quite large and can result in very different rankings.

On the other hand, due to substantially more sampling, the direct free energy method is expected to overcome this drawback. In this study, we have performed ESMACS calculations using all "bound" conformations (as defined so during ΔG_{direct} calculation) extracted from all "long" trajectories such that the ensemble averaging is performed across all replicas that sample a given binding site (denoted as "ESMACS-l"). Free energies so obtained are expected to be free from the dependence on starting structures and better correlated with ΔG_{direct}. This is evident in Figure 16 where $\Delta G_{ESMACS-s}$ as well as $\Delta G_{ESMACS-l}$ are compared against ΔG_{direct}. $\Delta G_{ESMACS-l}$ are consistently more negative than $\Delta G_{ESMACS-s}$ and have a higher correlation coefficient of 0.95. This is because all the different binding poses sampled during the long duration of simulations have been taken into account with appropriate weights.

4.2 Elucidation of an Allosteric Mechanism

Ligand RQN binds to the active binding site of the 3CLPro target protein whereas LZE binds to the allosteric binding site II[71]. In this study, we have performed simulations that contain both RQN and LZE ligands binding to the 3CLPro target at the same time. Therefore, we discuss the observed effect of the presence of LZE on the binding of RQN ligand by comparing the results from this system with

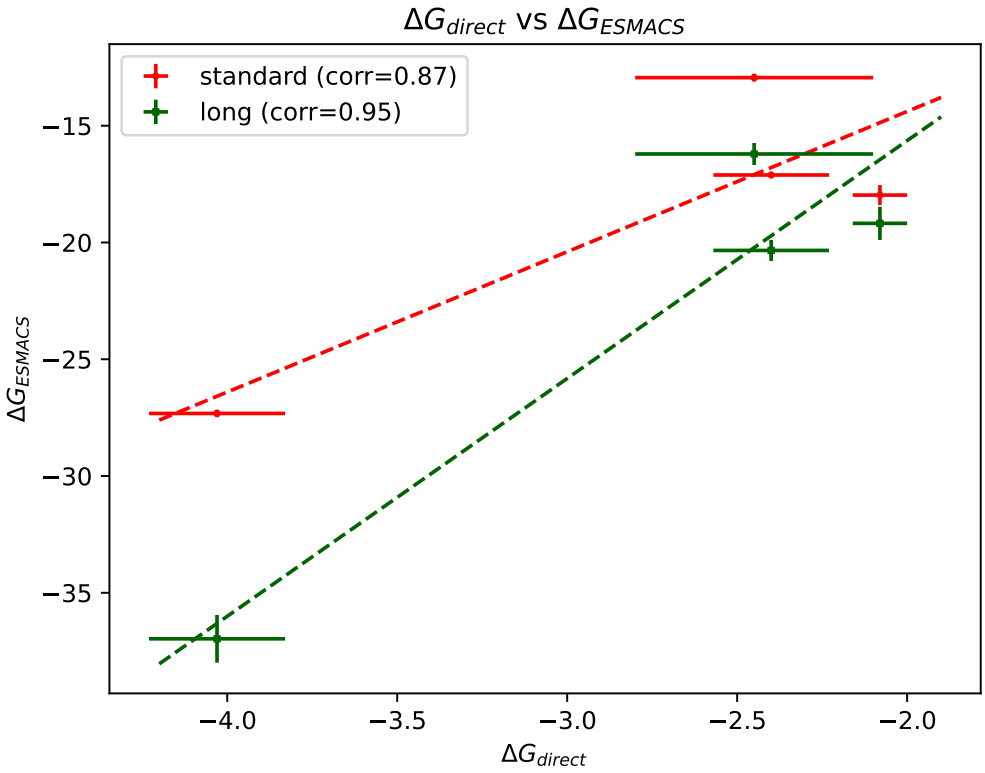

Figure 16: ΔG values obtained from different free energy protocols: ΔG_{direct} compared against ΔG_{ESMACS} using both the standard ESMACS protocol as well as that using bound conformations extracted from "long" trajectories. "corr" denotes the Pearson's correlation coefficient. Dashed lines denote the best fit lines for each plot. All values are in kcal/mol.

those from the system containing only RQN. First of all, the presence of LZE does not affect the value of ΔG_{direct} for RQN binding with 3CLPro. The respective values in the presence and absence of LZE are -3.04 ± 0.39 kcal/mol and -3.32 ± 0.23 which are statistically the same. However, the important thing to note here is that the respective spreads (difference between extremes) in ΔG values for these systems are 1.54 kcal/mol (ranging from -3.97 to -2.43 kcal/mol) and 2.64 kcal/mol (ranging from -4.56 to -1.92 kcal/mol) which are both much larger than the difference between their mean ΔG values. This indicates the importance of performing ensembles in order to obtain statistically robust and reliable conclusions. For instance, taking the opposite extremes of ΔG_{direct} values for both systems, we could have obtained differences of either -2.05 kcal/mol or 2.13 kcal/mol in the presence and absence of LZE, respectively, leading to diametrically opposite conclusions on its effect on the binding of RQN. But on performing ensemble simulations, we are able to state with confidence that no statistically significant effect has been observed.

Qualitatively, another important effect that has been observed is the emergence of a new binding site for RQN (denoted as C2), very close to the experimentally observed binding site (denoted as C), when binding to 3CLPro active site in the presence of LZE. Experimentally, it has been shown that the binding of LZE at allosteric site II displaces the loop 153-155 such that C_α atom of TYR154 moves 2.8 Å, accompanied by a conformational change of ASP153[71]. This loop is connected to loop 167-172 through a β-sheet strand 156-166 which is expected to cause a shift in the former as well. Figure 17 displays binding sites C (red) and C2 (blue) for RQN in the form of observed volume occupancy maps Binding sites C and C2 have loops 167-172 and 186-191 in common (shown in green). Loops 40-43 and 141-145 (shown in orange) are exclusive to site C, whereas loop 182-185 and residues 134-135 are exclusive to site C2. Therefore, the binding of LZE at allosteric site II brings about conformational changes to the active site and creates enough space to let RQN bind at a slightly different location, very close to the original site. It appears that such a change does not have any substantial impact on the binding interactions of RQN with the residues of site C, thereby not affecting its binding affinity. However, its sampling frequency is certainly affected such that, in the absence of LZE, it is sampled by 10 out of 12 "long" replicas, whereas in its presence, it is sampled only by 3 out of 12 "long" replicas. On the other hand, site C2 is sampled in 2 out of 12 "long" replicas (exclusively in the presence of LZE). A similar effect has been observed in case of LZE with RQN present such that a new close-by binding site (denoted as B2) is sampled along with the crystallographically determined binding site (denoted as B1). Table 2 and Figure 6 include both such binding sites (for both RQN and LZE in presence of each other) as experimental binding sites and display results accordingly.

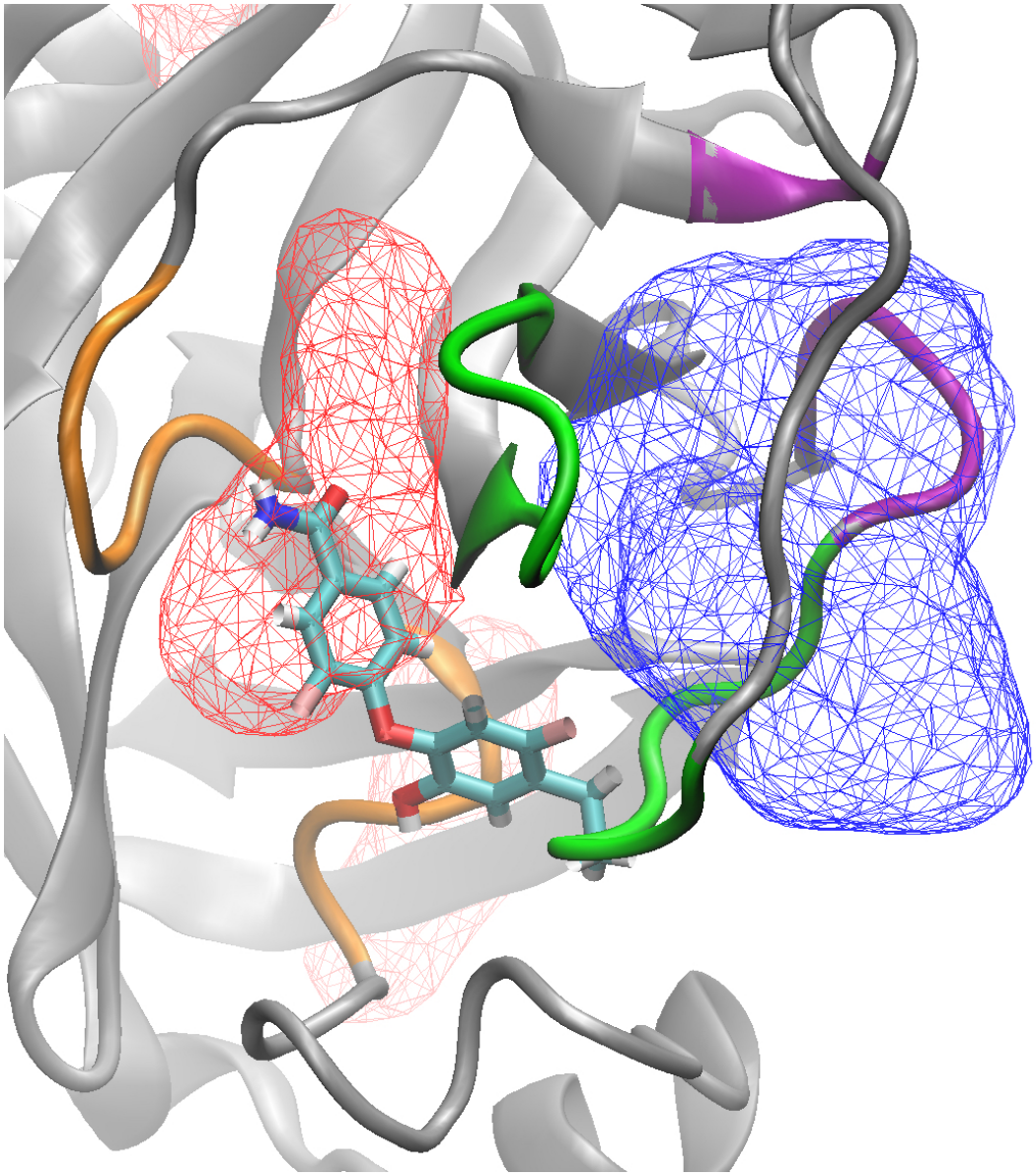

Figure 17: Effect of the presence of LZE on the binding of RQN. The experimental binding site (red) as well as the alternate binding site observed (blue) are shown in terms of volume occupancy maps using wideframe isosurfaces at isovalue 0.3. The crystallographically determined binding pose has also been shown in the "bonds" representation. Loops 167-172 and 186-191 (shown in green) are common to both binding sites. Loops 40-43 and 141-145 (shown in orange) are exclusive to the experimental binding site, whereas loop 182-185 and residues 134-135 are exclusive to the alternative site observed.

5 Conclusions

The current consensus in the field of molecular dynamics simulation is that increasing the length of a single simulation leads to improvement in the accuracy and precision of calculated expectation values [94–96]. On the basis of chaos theory, and the fact that the ergodic theorem cannot hold for molecular dynamics simulations on accessible timescales, we probed this assumption and provided direct evidence that individual trajectories do not suffice for deriving precise, reproducible and accurate results. We showed on the contrary that ensembles are essential for the calculation of statistically robust results, regardless of the length of simulation. On comparing the protein-ligand contact frequency distributions from ten or twelve independent 10 μs trajectories, 90% or more pairs of trajectories had significantly different distributions of ligand-protein residue interactions. We would like to emphasise that the principles and findings of this study are not just confined to ligand-protein systems and free energy calculations, but are more widely applicable to molecular dynamics in general and hence should be accounted for in all MD based applications regardless of the particular domain of interest [23].

To investigate the effect of this uncertainty on the value of a one dimensional macroscopic observable, we analysed the same set of trajectories in order to determine ligand binding free energies and their associated statistical distributions. The specific method which we used for ligand binding free energy calculations was taken from Pan et al. [27]. In their paper, the authors reported strong correlation to FEP calculations but poor correlation to experiment, stating this poor correlation may be attributable to force field inaccuracies. In the present study, we demonstrated that separate trajectories lead to the computation of completely different results, differing by up to 7.26 kcal/mol. Our study conclusively demonstrates that binding free energies from individual simulations are inherently non-precise, non-reproducible and do not yield chemical accuracy ($\pm$ 1 kcal/mol). Clearly, long-timescale trajectories probe an insufficient number of microstates to effectively sample the phase space. In turn, the lack of agreement with experiment should not necessarily be attributed to force-field inaccuracies. This is a paramount example of the importance of taking aleatoric uncertainty fully into account. This principle holds for the expectation value of any other dynamical observable obtained via "long-timescale" simulation since it depends on the nature of the probability distribution (or invariant measure) which is intrinsic to the system of interest.

In addition, by executing both the long-timescale and splitting protocols we have provided insight into the utility of adaptive sampling protocols. With respect to the length of simulations, it is clear that the merit of running a long simulation changes as a function of the timescale of events of interest. In the

case of the systems studied here, no long timescale events (e.g. large-scale domain rearrangements) need to occur for ligand binding to be possible. As a result, a simple adaptive sampling protocol was able to successfully identify all of the sites identified by the long timescale protocol albeit with significantly less wall time required (1 day 7 hours for adaptive sampling as compared to 43 days 2 hours for 10 microseconds of simulation for the ADRP system).

Beyond these implications, the findings in this work also show how ensemble based computational protocols can be used to inform the process of drug discovery. For instance, with respect to ADRP, 4 binding sites that tofacitinib can bind to were identified within both the long-timescale and splitting protocols. From our binding free energy analysis, we identified that tofacitinib binds to the crystallographically determined binding site with the greatest affinity out of each of ADRP the binding sites. This indicates that, in practice, tofacitinib would act as a competitive inhibitor of ADRP. Similarly, various binding sites of interest were identified for other ligand-protein complexes studied with similar conclusions made. In addition, the discovery of non-crystallographically resolved binding sites is of great interest for a future study which would aim to elucidate whether any of these binding sites can propagate allosteric effects to the substrate binding site. This would provide a novel mechanism by which to target the protein and induce anti-viral effects. Finally, we compared the "direct" free energy method with ESMACS and discussed various scenarios where each method has an advantage or limitation. ESMACS is very efficient in ranking ligands based on their binding interactions with much less computational cost as compared to the direct binding affinity method. However, it is subject to the availability of a stable binding pose as the starting structure, in the absence of which long simulations do a better job. We hope that this will help others working in this domain to choose an appropriate free energy method for their purposes.

Finally, the use of ensemble methods enabled us to discover the allosteric mechanism through which the binding of a ligand at the substrate binding site of 3CLPro is affected by binding of another ligand at an experimentally known allosteric binding site. We showed that the two binding sites are connected via a β-sheet strand that causes distortion to the cavity of the substrate binding site relative to its conformation in the absence of such an allosteric effect.

Author Contributions

APB and PVC conceptualised the study. AH, APB and AP prepared models and performed simulations. AH and APB conducted data analyses. AH, APB and PVC wrote the manuscript, while AH, APB, AP and PVC participated in the manuscript editing. PVC supervised the work and acquired computational

resources and funding for the study.

Conflicts of interest

There are no conflicts to declare.

Supporting Information

Figures displaying contact frequency distributions, KS statistics, p-boxes and cumulative density functions as well as comparisons of contact frequency distributions from long simulations and splitting protocols have been included in the Supporting Information for all systems that were not accommodated in the main text. All input structure and parameter files are available on a public github repository at `https://github.com/UCL-CCS/LongTimescaleStudy`.

Acknowledgements

The authors would like to acknowledge funding support from (i) the UK EPSRC for the UK High-End Computing Consortium (EP/R029598/1) and the Software Environment for Actionable & VVUQ-evaluated Exascale Applications (SEAVEA) grant (EP/W007762/1); (ii) A 2021 DOE INCITE award of 125,000 Summit node hours and 100,000 Theta node hours under the 'COMPBIO' project; (iii) The Hartree Centre for unlimited CPU and GPU core hours on the Scafell Pike supercomputer; (iv) the European Union's Horizon 2020 Research and Innovation Programme under grant agreement 823712 (CompBioMed2, compbiomed.eu), and funding from the UCL Provost. We acknowledge the United States Department of Energy (DOE) and Oak Ridge National Laboratory (ORNL) for providing access to and core hours on Summit (https://www.olcf.ornl.gov/summit/). We give particular thanks to Benjamin Aguirre Hernandez and Bronson Messer at ORNL for providing invaluable support, advice and special access to resources without which this work would not have been possible. Furthermore, we are grateful to all of the staff at ORNL and at the Hartree Centre for their generous support. Simulations were carried out with NAMD 2 (`https://www.ks.uiuc.edu/Research/namd/`) and OpenMM (`https://openmm.org/`).

References

[1] N. Berdigaliyev and M. Aljofan, *Future Medicinal Chemistry*, 2020, **12**, 939–947.

[2] A. A. Saadi, D. Alfe, Y. Babuji, A. Bhati, B. Blaiszik, T. Brettin, K. Chard, R. Chard, P. Coveney,

A. Trifan, A. Brace, A. Clyde, I. Foster, T. Gibbs, S. Jha, K. Keipert, T. Kurth, D. Kranzlmüller, H. Lee, Z. Li, H. Ma, A. Merzky, G. Mathias, A. Partin, J. Yin, A. Ramanathan, A. Shah, A. Stern, R. Stevens, L. Tan, M. Titov, A. Tsaris, M. Turilli, H. Van Dam, S. Wan and D. Wifling, 50th International Conference on Parallel Processing, 2021, pp. 1–12.

[3] A. P. Bhati, S. Wan, D. Alfè, A. R. Clyde, M. Bode, L. Tan, M. Titov, A. Merzky, M. Turilli, S. Jha, R. R. Highfield, W. Rocchia, N. Scafuri, S. Succi, D. Kranzmüller, G. Mathias, D. Wifling, Y. Donon, A. Di Meglio, S. Vallecorsa, H. Ma, A. Trifan, A. Ramanathan, T. Brettin, A. Partin, F. Xia, X. Duan, R. Stevens and P. V. Coveney, *J. R. Soc. Interface Focus*, 2021, **11**, 20210018.

[4] H. Lee, A. Merzky, L. Tan, M. Titov, M. Turilli, D. Alfe, A. Bhati, A. Brace, A. Clyde, P. Coveney *et al.*, Proceedings of the Platform for Advanced Scientific Computing Conference, 2021, pp. 1–13.

[5] D. B. Korlepara, C. Vasavi, S. Jeurkar, P. K. Pal, S. Roy, S. Mehta, S. Sharma, V. Kumar, C. Muvva, B. Sridharan *et al.*, *Scientific data*, 2022, **9**, 1–10.

[6] I. Buch, T. Giorgino and G. D. Fabritiis, *Proceedings of the National Academy of Sciences*, 2011, **108**, 10184–10189.

[7] R. O. Dror, A. C. Pan, D. H. Arlow, D. W. Borhani, P. Maragakis, Y. Shan, H. Xu and D. E. Shaw, *Proceedings of the National Academy of Sciences*, 2011, **108**, 13118–13123.

[8] Y. Shan, V. P. Mysore, A. E. Leffler, E. T. Kim, S. Sagawa and D. E. Shaw, *bioRxiv*, 2021, 2021.03.31.437917.

[9] Y. Shan, E. T. Kim, M. P. Eastwood, R. O. Dror, M. A. Seeliger and D. E. Shaw, *Journal of the American Chemical Society*, 2011, **133**, 9181–9183.

[10] D.-A. Silva, G. R. Bowman, A. Sosa-Peinado and X. Huang, *PLoS Computational Biology*, 2011, **7**, e1002054.

[11] S. Gu, D.-A. Silva, L. Meng, A. Yue and X. Huang, *PLOS Computational Biology*, 2014, **10**, e1003767.

[12] N. Ahalawat and J. Mondal, *Journal of the American Chemical Society*, 2018, **140**, 17743–17752.

[13] R. M. Betz and R. O. Dror, *Journal of Chemical Theory and Computation*, 2019, **15**, 2053–2063.

[14] S. Bikkina, A. P. Bhati, S. Padhi and U. D. Priyakumar, *Journal of Chemical Sciences*, 2017, **129**, 405–414.

[15] D. E. Shaw, J. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton *et al.*, SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2014, pp. 41–53.

[16] D. E. Shaw, P. J. Adams, A. Azaria, J. A. Bank, B. Batson, A. Bell, M. Bergdorf, J. Bhatt, J. A. Butts, T. Correia *et al.*, SC'21: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2021, pp. 1–11.

[17] D. Groen, A. P. Bhati, J. Suter, J. Hetherington, S. J. Zasada and P. V. Coveney, *Computer Physics Communications*, 2016, **207**, 375–385.

[18] M. K. Bieniek, A. P. Bhati, S. Wan and P. V. Coveney, *Journal of chemical theory and computation*, 2021, **17**, 1250–1265.

[19] P. V. Coveney and S. Wan, *Physical Chemistry Chemical Physics*, 2016, **18**, 30236–30240.

[20] A. P. Bhati, S. Wan, D. W. Wright and P. V. Coveney, *Journal of Chemical Theory and Computation*, 2017, **13**, 210–222.

[21] A. P. Bhati, S. Wan, Y. Hu, B. Sherborne and P. V. Coveney, *Journal of Chemical Theory and Computation*, 2018, **14**, 2867–2880.

[22] D. W. Wright, B. A. Hall, O. A. Kenway, S. Jha and P. V. Coveney, *Journal of Chemical Theory and Computation*, 2014, **10**, 1228–1241.

[23] S. Wan, R. C. Sinclair and P. V. Coveney, *Philosophical Transactions of the Royal Society A*, 2021, **379**, 20200082.

[24] M. Vassaux, S. Wan, W. Edeling and P. V. Coveney, *Journal of Chemical Theory and Computation*, 2021, **17**, 5187–5197.

[25] V. S. Pande, K. Beauchamp and G. R. Bowman, *Methods*, 2010, **52**, 99 – 105.

[26] G. R. Bowman, X. Huang and V. S. Pande, *Methods*, 2009, **49**, 197 – 201.

[27] A. C. Pan, H. Xu, T. Palpant and D. E. Shaw, *Journal of Chemical Theory and Computation*, 2017, **13**, 3372–3377.

[28] N. Plattner and F. Noé, *Nature Communications*, 2015, **6**, 7653.

[29] F. Paul, C. Wehmeyer, E. T. Abualrous, H. Wu, M. D. Crabtree, J. Schöneberg, J. Clarke, C. Freund, T. R. Weikl and F. Noé, *Nature Communications*, 2017, **8**, 1095.

[30] R. O. Dror, H. F. Green, C. Valant, D. W. Borhani, J. R. Valcourt, A. C. Pan, D. H. Arlow, M. Canals, J. R. Lane, R. Rahmani, J. B. Baell, P. M. Sexton, A. Christopoulos and D. E. Shaw, *Nature*, 2013, **503**, 295–299.

[31] S. Doerr and G. De Fabritiis, *Journal of Chemical Theory and Computation*, 2014, **10**, 2064–2069.

[32] K. J. Kohlhoff, D. Shukla, M. Lawrenz, G. R. Bowman, D. E. Konerding, D. Belov, R. B. Altman and V. S. Pande, *Nature Chemistry*, 2014, **6**, 15–21.

[33] G. R. Bowman, K. A. Beauchamp, G. Boxer and V. S. Pande, *The Journal of Chemical Physics*, 2009, **131**, 124101.

[34] F. Noé and S. Fischer, *Current Opinion in Structural Biology*, 2008, **18**, 154 – 162.

[35] S. Wan, A. P. Bhati, S. J. Zasada and P. V. Coveney, *J. R. Soc. Interface Focus*, 2020, **10**, 20200007.

[36] S. Wan, A. P. Bhati, S. J. Zasada, I. Wall, D. Green, P. Bamborough and P. V. Coveney, *Journal of Chemical Theory and Computation*, 2017, **13**, 784–795.

[37] S. Wan, A. P. Bhati, S. Skerratt, K. Omoto, V. Shanmugasundaram, S. K. Bagal and P. V. Coveney, *Journal of Chemical Information and Modeling*, 2017, **57**, 897–909.

[38] S. Wan, A. P. Bhati, A. D. Wade, D. Alfè and P. V. Coveney, *Mol. Syst. Des. Eng.*, 2022, **7**, 123–131.

[39] A. F. Voter, *Phys. Rev. B*, 1998, **57**, R13985–R13988.

[40] M. R. Shirts and V. S. Pande, *Phys. Rev. Lett.*, 2001, **86**, 4983–4987.

[41] B. Zagrovic, E. J. Sorin and V. Pande, *Journal of Molecular Biology*, 2001, **313**, 151 – 169.

[42] C. D. Snow, H. Nguyen, V. S. Pande and M. Gruebele, *Nature*, 2002, **420**, 102–106.

[43] W. C. Swope, J. W. Pitera and F. Suits, *The Journal of Physical Chemistry B*, 2004, **108**, 6571–6581.

[44] D. L. Ensign, P. M. Kasson and V. S. Pande, *Journal of Molecular Biology*, 2007, **374**, 806 – 816.

[45] G. Jayachandran, V. Vishal and V. S. Pande, *The Journal of Chemical Physics*, 2006, **124**, 164902.

[46] J. D. Chodera, W. C. Swope, J. W. Pitera and K. A. Dill, *Multiscale Modeling & Simulation*, 2006, **5**, 1214–1226.

[47] N. Singhal and V. S. Pande, *The Journal of Chemical Physics*, 2005, **123**, 204909.

[48] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill and W. C. Swope, *The Journal of Chemical Physics*, 2007, **126**, 155101.

[49] G. Huber and S. Kim, *Biophysical Journal*, 1996, **70**, 97 – 110.

[50] D. M. Zuckerman and L. T. Chong, *Annual Review of Biophysics*, 2017, **46**, 43–57.

[51] A. Dickson, *Biophysical Journal*, 2018, **115**, 1707 – 1719.

[52] D. Bhatt, B. W. Zhang and D. M. Zuckerman, *The Journal of Chemical Physics*, 2010, **133**, 014110.

[53] D. Bhatt and I. Bahar, *The Journal of Chemical Physics*, 2012, **137**, 104101.

[54] J. L. Adelman and M. Grabe, *The Journal of Chemical Physics*, 2013, **138**, 044105.

[55] E. Suárez, S. Lettieri, M. C. Zwier, C. A. Stringer, S. R. Subramanian, L. T. Chong and D. M. Zuckerman, *Journal of Chemical Theory and Computation*, 2014, **10**, 2658–2667.

[56] M. C. Zwier, J. L. Adelman, J. W. Kaus, A. J. Pratt, K. F. Wong, N. B. Rego, E. Suárez, S. Lettieri, D. W. Wang, M. Grabe, D. M. Zuckerman and L. T. Chong, *Journal of Chemical Theory and Computation*, 2015, **11**, 800–809.

[57] A. Dickson and C. L. Brooks, *The Journal of Physical Chemistry B*, 2014, **118**, 3532–3542.

[58] B. Abdul-Wahid, H. Feng, D. Rajan, R. Costaouec, E. Darve, D. Thain and J. A. Izaguirre, *Journal of Chemical Information and Modeling*, 2014, **54**, 3033–3043.

[59] P. Glasserman, P. Heidelberger, P. Shahabuddin and T. Zajic, Proceedings Winter Simulation Conference, 1996, pp. 302–308.

[60] P. Glasserman, P. Heidelberger, P. Shahabuddin and T. Zajic, *Operations Research*, 1999, **47**, 585–600.

[61] F. Cérou and A. Guyader, *Stochastic Analysis and Applications*, 2007, **25**, 417–443.

[62] I. Teo, C. G. Mayne, K. Schulten and T. Lelièvre, *Journal of Chemical Theory and Computation*, 2016, **12**, 2983–2989.

[63] F. Cérou, A. Guyader and M. Rousset, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 2019, **29**, 043108.

[64] H. Fukunishi, O. Watanabe and S. Takada, *The Journal of Chemical Physics*, 2002, **116**, 9058–9067.

[65] L. Wang, R. A. Friesner and B. J. Berne, *The Journal of Physical Chemistry B*, 2011, **115**, 9431–9438.

[66] A. P. Bhati, S. Wan and P. V. Coveney, *Journal of Chemical Theory and Computation*, 2019, **15**, 1265–1277.

[67] K. Michalska, Y. Kim, R. Jedrzejczak, N. I. Maltseva, L. Stols, M. Endres and A. Joachimiak, *IUCrJ*, 2020, **7**, 814–824.

[68] Z. Fu, B. Huang, J. Tang, S. Liu, M. Liu, Y. Ye, Z. Liu, Y. Xiong, W. Zhu, D. Cao, J. Li, X. Niu, H. Zhou, Y. J. Zhao, G. Zhang and H. Huang, *Nature Communications*, 2021, **12**, 1–12.

[69] H.-x. Su, S. Yao, W.-f. Zhao, M.-j. Li, J. Liu, W.-j. Shang, H. Xie, C.-q. Ke, H.-c. Hu, M.-n. Gao, K.-q. Yu, H. Liu, J.-s. Shen, W. Tang, L.-k. Zhang, G.-f. Xiao, L. Ni, D.-w. Wang, J.-p. Zuo, H.-l. Jiang, F. Bai, Y. Wu, Y. Ye and Y.-c. Xu, *Acta Pharmacologica Sinica*, 2020, **41**, 1167–1177.

[70] P. V. Coveney, D. Groen and A. G. Hoekstra, *Reliability and reproducibility in computational science: implementing validation, verification and uncertainty quantification in silico*, 2021.

[71] S. Günther, A. Meents, P. Y. Reinke, Y. Fernández-García, J. Lieske, T. J. Lane, H. M. Ginn, F. H. Koua, C. Ehrt, W. Ewert, D. Oberthuer *et al.*, *Science*, 2021, **372**, 642–646.

[72] S. B. Cohen, Y. Tanaka, X. Mariette, J. R. Curtis, E. B. Lee, P. Nash, K. L. Winthrop, C. Charles-Schoeman, K. Thirunavukkarasu, R. DeMasi, J. Geier, K. Kwok, L. Wang, R. Riese and J. Wollenhaupt, *Annals of the Rheumatic Diseases*, 2017, **76**, 1253–1262.

[73] E. J. Kucharz, M. Stajszczyk, A. Kotulska-Kucharz, B. Batko, M. Brzosko, S. Jeka, P. Leszczyński, M. Majdan, M. Olesińska, W. Samborski and P. Wiland, *Reumatologia/Rheumatology*, 2018, **56**, 203–211.

[74] D. W. Wright, S. Wan, C. Meyer, H. van Vlijmen, G. Tresadern and P. V. Coveney, *Scientific Reports*, 2019, **9**, 6017.

[75] S. Wan, A. Potterton, F. S. Husseini, D. W. Wright, A. Heifetz, M. Malawski, A. Townsend-Nicholson and P. V. Coveney, *J. R. Soc. Interface Focus*, 2020, **10**, 20190128.

[76] M. K. Bieniek, A. P. Bhati, S. Wan and P. V. Coveney, *Journal of Chemical Theory and Computation*, 2021, **17**, 1250–1265.

[77] M. V. Shapovalov and R. L. Dunbrack Jr, *Structure*, 2011, **19**, 844–858.

[78] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin, *Journal of Computational Chemistry*, 2004, **25**, 1605–1612.

[79] L. Kalé, R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan and K. Schulten, *Journal of Computational Physics*, 1999, **151**, 283–312.

[80] Y. Shan, V. P. Mysore, A. E. Leffler, E. T. Kim, S. Sagawa and D. E. Shaw, *PLOS Computational Biology*, 2022, **18**, 1–20.

[81] P. Robustelli, A. Ibanez-de Opakua, C. Campbell-Bezat, F. Giordanetto, S. Becker, M. Zweckstetter, A. C. Pan and D. E. Shaw, *Journal of the American Chemical Society*, 2022, **144**, 2501–2510.

[82] J. C. Phillips, D. J. Hardy, J. D. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M. C. R. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot and E. Tajkhorshid, *The Journal of Chemical Physics*, 2020, **153**, 044130.

[83] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks and V. S. Pande, *PLoS Computational Biology*, 2017, **13**, e1005659.

[84] A. Venkatakrishnan, R. Fonseca, A. K. Ma, S. A. Hollingsworth, A. Chemparathy, D. Hilger, A. J. Kooistra, R. Ahmari, M. M. Babu, B. K. Kobilka and R. O. Dror, *bioRxiv*, 2019, 840694.

[85] D. H. De Jong, L. V. Schäfer, A. H. De Vries, S. J. Marrink, H. J. Berendsen and H. Grubmüller, *Journal of Computational Chemistry*, 2011, **32**, 1919–1928.

[86] P. A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D. A. Case and T. E. Cheatham, *Accounts of Chemical Research*, 2000, **33**, 889–897.

[87] D.A. Case, K. Belfon, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, A. Kovalenko, R. Krasny, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, V. Man, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, A. Onufriev, F.Pan, S. Pantano, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, N.R.Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, L. Wilson, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York and P.A. Kollman (2020), AMBER 2020, University of California, San Francisco.

[88] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors, *Nature Methods*, 2020, **17**, 261–272.

[89] W. L. Oberkampf and C. J. Roy, *Verification and Validation in Scientific Computing*, Cambridge University Press, 2010.

[90] A. P. Bhati and P. V. Coveney, *Journal of Chemical Theory and Computation*, 2022, **18**, 2687–2702.

[91] A. D. Wade, A. P. Bhati, S. Wan and P. V. Coveney, *Journal of Chemical Theory and Computation*, 2022, **18**, 3972–3987.

[92] S. Wan, A. P. Bhati, D. W. Wright, I. D. Wall, A. P. Graves, D. Green and P. V. Coveney, *Journal of Chemical Information and Modeling*, 2022, **62**, 2561–2570.

[93] S. Wan, A. P. Bhati, D. W. Wright, A. D. Wade, G. Tresadern, H. van Vlijmen and P. V. Coveney, *Scientific Reports*, 2022, **12**, 10433.

[94] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Elsevier, 2001, pp. 15–17.

[95] J. Haile, *Molecular Dynamics Simulation: Elementary Methods*, Wiley, 1997, pp. 15–16.

[96] R. D. Skeel, *SIAM Journal on Scientific Computing*, 2009, **31**, 1363–1378.