Sensitivity of Binding Free Energy Calculations to Initial Protein Crystal Structure

Miroslav Suruzhon,[†] Michael S. Bodnarchuk,[‡] Antonella Ciancetta,[¶] Russell Viner,[§] Ian D. Wall,[‖] and Jonathan W. Essex[*,†]

[†]*School of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom*

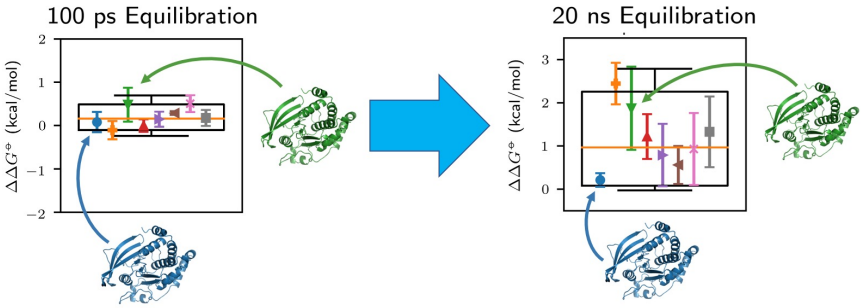
[‡]*Computational Chemistry, R&D Oncology, AstraZeneca, Cambridge CB4 0WG, United Kingdom*

[¶]*Sygnature Discovery Limited, BioCity, Pennyfoot St, Nottingham NG1 1GR, United Kingdom*

[§]*Syngenta, Jealott's Hill International Research Centre, Bracknell RG42 6EY, United Kingdom*

[‖]*GSK Medicines Research Centre, Gunnels Wood Road, Stevenage SG1 2NY, United Kingdom*

E-mail: J.W.Essex@soton.ac.uk



Abstract

Binding free energy calculations using alchemical free energy (AFE) methods are widely considered to be the most rigorous tool in the computational drug discovery arsenal. Despite this, the calculations suffer from accuracy, precision and reproducibility issues. In this publication, we perform a high-throughput study of more than a thousand AFE calculations, utilizing over 220 μs of total sampling time, on three different protein systems to investigate the impact of the initial crystal structure on the resulting binding free energy values. We also consider the influence of equilibration time and discover that the initial crystal structure can have a significant effect on free energy values obtained at short timescales that can manifest itself as a free energy difference of more than 1 kcal/mol. At longer timescales, these differences are largely overtaken by important rare events, such as torsional ligand motions, typically resulting in a much higher uncertainty in the obtained values. This work emphasizes the importance of rare event sampling and long-timescale dynamics in free energy calculations even for routinely performed alchemical perturbations. We conclude that an optimal protocol should not only concentrate the computational resources on achieving convergence in the alchemical coupling parameter (λ) space, but also on longer simulations and multiple repeats.

1 Introduction

There are two overarching topics of research in computational biophysics: improvement of model accuracy (solved by developing new Hamiltonians) and increasing the precision of model evaluation (solved by more efficient enhanced sampling methods). However, there is a third problem which is also relevant to all areas of science – reproducibility. The exploration of this issue has been gathering increased attention from computational chemists in recent years, not least because it is a requirement and an expectation from any method nearing operational maturity.

An example is alchemical free energy methods (AFE), which have amassed increasing popularity not only as a topic of academic research but also as a tool in commercial drug discovery[1–3] since their conception. However, the nature of these methods requires the scientist to make various choices in the initial conditions and/or the "hyperparameters" of the underlying models. These choices are occasionally obvious, sometimes conventional and often completely arbitrary. Robustness of the predictive model to these choices would imply consistent results between individual scientists, research groups, software packages and therefore facilitate automatability – a highly desirable quality in the age of high-performance high-throughput computing.

Recent efforts have investigated the influence of the following factors on free energy calculations: the force field;[4–8] the type of enhanced sampling method;[8–11] the sampling time;[7,9,12,13] the barostat;[10] the free energy calculation method;[10,14,15] the free energy analysis method;[16] the software package;[7,8,10,17,18] the alchemical interpolation protocol;[19] the charge perturbation method;[20] the protonation and tautomeric states;[21] the location of water molecules in the binding site;[22–26] the ligand binding mode.[26,27] In addition, it has been suggested that system preparation is a crucial part of the free energy calculation process.[2,7] Directly relevant to our work are previous results showing free energy dependence on the protein conformation[28] and on the ligand alignment based on the initial protein crystal structure used.[7,25]

A number of the above studies have found concerning reproducibility issues. Rizzi et al.[10] report differences of between 0.3 and 1.0 kcal/mol between different equally valid methods using the same force field parameters. Similarly, Loeffler et al.[18] show that even for simple solvation free energy calculations, reproducibility between software packages is only maintained to 0.2 kcal/mol. Even more significant is the impact of initial binding site water molecules on the free energy estimate, in some cases reaching discrepancies of over 5 kcal/mol,[25] meaning that assuming a particular binding mode without performing enhanced sampling can be detrimental to the free energy calculation. Pérez-Benito et al.[7] observe sim-

ilar behaviour when comparing free energy results obtained from different crystal structures native to either of the perturbed ligands with some discrepancies surpassing 5 kcal/mol. Even though this behaviour is mitigated to some extent by increased sampling, some significant outliers can still persist, meaning that the initial system setup can be a non-trivial process and warrants further investigation.

In this work we examine the effect of the initial protein crystal structure on a range of ligand-ligand perturbations systematically and in depth. The crystal structures have been chosen so that they vary in their resolution, year of deposition, bound ligands and research groups. Ligand perturbations have generally been kept simple, since we are mostly interested in calculations that appear "converged" at short timescales. We study three distinct protein systems – dihydrofolate reductase (DHFR), protein tyrosine phosphatase 1B (PTP1B) and factor Xa (FXa). These test systems have been chosen so that the following commonly encountered features are covered at least once: cofactors, auxiliary ions, disulfide bonds, missing residues and atoms, multiple chains. In addition, these proteins are relatively small and shown in previous computational studies[29–32] to compare favourably to experiment, meaning that major force field and efficiency issues are not expected. We perform a high-throughput study on the effect of different crystal structures on the relative binding free energy values with and without an extra 20 ns equilibration. This will allow us to verify the increasingly common notion that multiple short simulations are preferable to a single long one.[33–35] In all cases we also analyze cycle closure errors and compare to experiment whenever possible. To the best of our knowledge, this is the most extensive study to date which investigates the dependence of binding free energies on the initial structure.

2 Methods

2.1 System Preparation

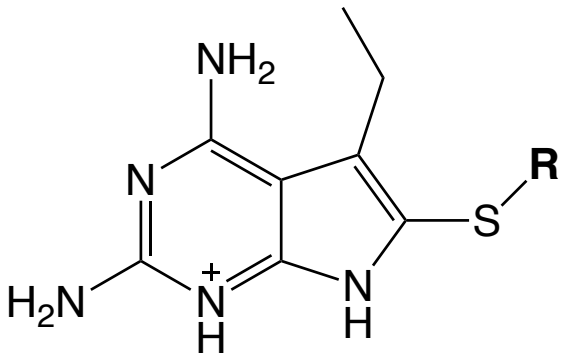
All system preparation in this study was performed using ProtoCaller.[36] The following X-ray crystal structures were used in this study: 1OHJ,[37] 2W3M,[38] 3GHW,[39] 4DDR,[40] 4M6J,[41] 5HPB,[42] 6A7E[43] and 6DAV[44] for DHFR; 1BZJ,[45] 1NWE,[46] 2AZR,[47] 2F71,[48] 2H4K,[49] 2NTA,[50] 2QBP[51] and 2ZN7[52] for PTP1B; 1EZQ,[53] 1KSN,[54] 1LQD,[55] 1NFW,[56] 2CJI,[57] 2J38,[58] 2J95[59] and 4Y71[60] for FXa. All protein crystal structures were obtained from the Protein Data Bank[61] (PDB). Some relevant metrics by which these crystal structures differ are shown in the Supporting Information (Tables S1 to S3). Most importantly, all of the above structures have a root-mean-suare deviation (RMSD) of less than 1 Å after alignment to a reference structure (described later) and only one of them (4M6J) has an RMSD larger than 0.5 Å, indicating that the initial structures used can all be considered very similar to one another. In addition, some of the structures (4M6J, 1NWE, 2NTA, 1EZQ, 1LQD, 1NFW, 2J38, 4Y71) exhibit slight differences in their reported protein sequence compared to the others and these differences were kept. None of these differences were near the binding site (within 1.2 nm of the ligand center of mass).

Where applicable, non-terminal missing residues were added using Modeller[62] and missing atoms were modelled using PDB2PQR.[63] All crystal structures were protonated with PDB2PQR, where all histidine residues were arbitrarily forced to the ϵ tautomer for FXa for consistency, due to the presence of multiple histidines near the binding site. No equivalent modifications to tautomer preference were made for DHFR and PTP1B, in some cases resulting in histidine tautomer differences across structures distal from the binding site (more than 1.2 nm away from the ligand center of mass). All amino acids were assigned their default expected protonation states at pH = 7. All crystal structure waters were retained. In two of the DHFR structures, there were two copies of the protein in the asymmetric unit and in these cases only the first chain of the PDB file was used in the subsequent simulations.

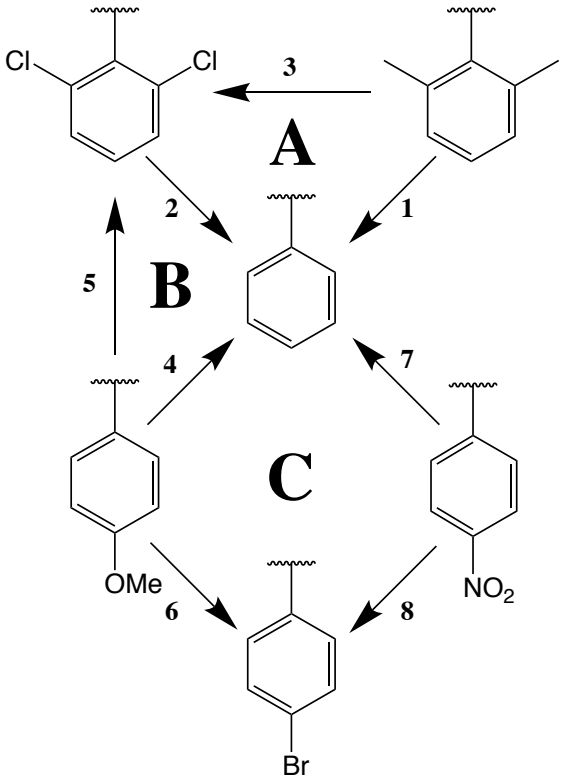
For each crystal structure we performed the following number of ligand-ligand perturbations: 8 for DHFR, 9 for PTP1B and 9 for FXa. The ligand scaffolds and thermodynamic cycles can be seen in Figure 1. Some of the ligands have been forced into an unnatural protonation state so that no charge perturbation was needed. Such ligands have only been treated as intermediates and have not been compared to experiment. The crystal structures whose native ligands most closely matched the scaffolds of our ligands of interest were used for alignment: 5HPB, 2QBP and 1LQD for DHFR, PTP1B and FXa, respectively. This initial ligand conformation was enforced across all other crystal structures based on protein-protein backbone alignment using PyMOL[64] and it was also used during the ligand parametrization stage in order to obtain the same partial charges for the same ligand across different crystal structures. Ligand protonation states were determined manually and were in most cases in agreement with Open Babel.[65] The most notable difference is the addition of an extra proton to the DHFR ligand at one of the pyrimidine nitrogen atoms (Figure 2), which has been previously suggested[66] to take part in ligand-carboxylate interactions involving a closely related ligand, methotrexate.

In this work we used the AMBER force field. All amino acid residues were parametrized with ff14SB,[67] and GAFF2[68] was used for all ligands. The TIP3P water model[69] and its associated calcium parameters were also used alongside the additional $NADP^+$ and NADPH parameters,[70] available online.[71] All ligand charges were parametrized using AM1-BCC.[72,73]

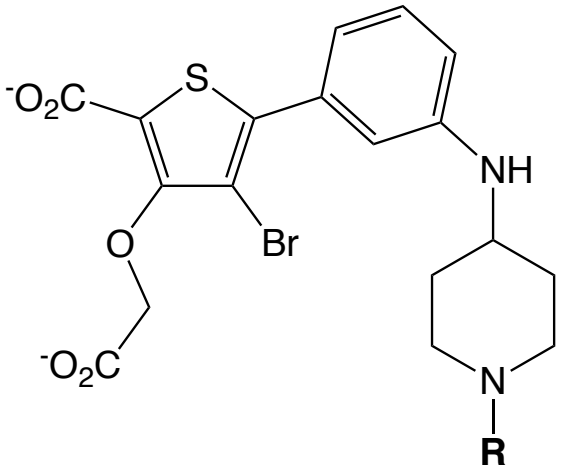
All ligand perturbations were performed using a single-topology mapping along an alchemical coordinate (λ) and alignment to the reference ligands was performed using ProtoCaller's default protocol by constraining the positions of the maximum common substructure atoms to the reference ones. Whenever applicable, all perturbations were in the direction of atom annihilation. All dummy atom equilibrium distances were scaled to half of their initial values in an attempt to achieve better phase space overlap between λ windows. All systems were solvated in cubic periodic boxes with a length of 7 nm in the bound leg and 4 nm in the solvated leg. NaCl was also added with an ionic concentration of ~ 0.154 M.



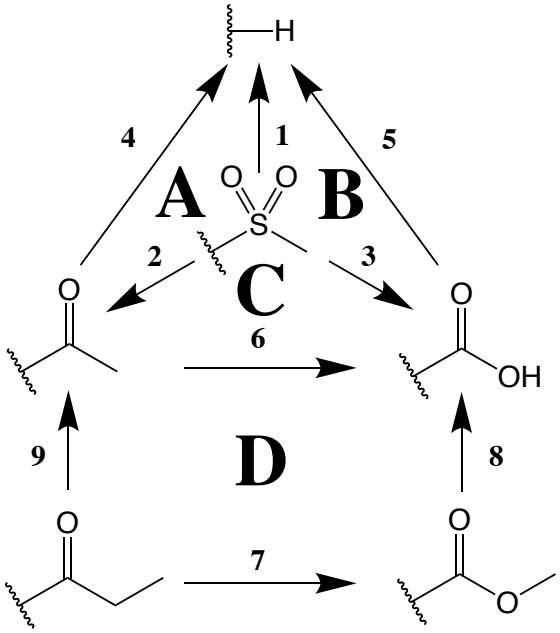
(a) DHFR ligand scaffold



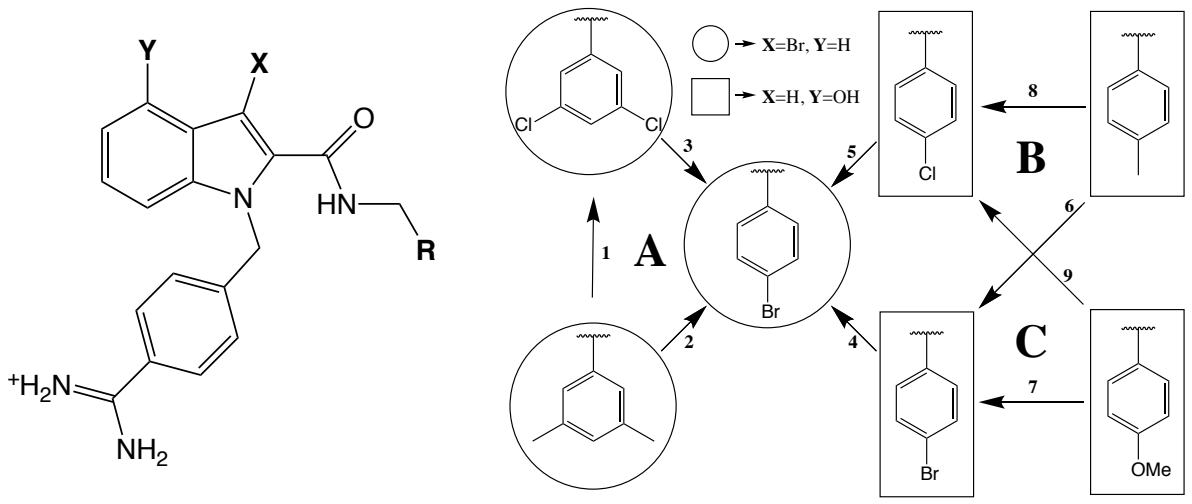
(b) DHFR perturbations



(c) PTP1B ligand scaffold



(d) PTP1B perturbations



(e) FXa ligand scaffold (f) FXa perturbations

Figure 1: Ligand scaffolds and perturbations for all three systems. The perturbed ligand pairs are denoted with numbers and thermodynamic cycles are labelled with letters. In all cases perturbations of **R** are shown. In Figure 1f, the circular ligands are substituted with **X**=Br and **Y**=H, and the rectangular ligands contain **X**=H and **Y**=OH.

2.2 Simulation

All simulations were performed in GROMACS 2018.4.[74] The perturbations were carried out over 40 λ windows with 10 equally-spaced perturbations of the electrostatic terms followed by simultaneous scaling of the Lennard-Jones (LJ) and bonded terms during the other 30 windows. The latter λ windows were also equally spaced (rounded to two significant figures), except for the final values, which were more closely spaced in an attempt to increase phase space overlap: 0.95, 0.97, 0.98, 0.99, 0.999 and 1. All interaction parameters were scaled linearly with respect to λ, except for the LJ interactions, which were perturbed using a soft-core potential[75] with a parameter value $\alpha = 0.5$.

Each λ window involved 25,000-step steepest descent minimization, 50 ps of NVT equilibration followed by 50 ps of NPT equilibration using a 1 fs timestep and 4 ns NPT production with a 2 fs timestep. In all cases the calculations were repeated after an additional initial 20 ns NPT equilibration at $\lambda = 0$, with this coordinate set being used to prepare simulations at all λ values. Both the 100 ps and the 20 ns protocols were repeated in triplicates, where the only difference was the pseudorandom number seed used for velocity initialization from

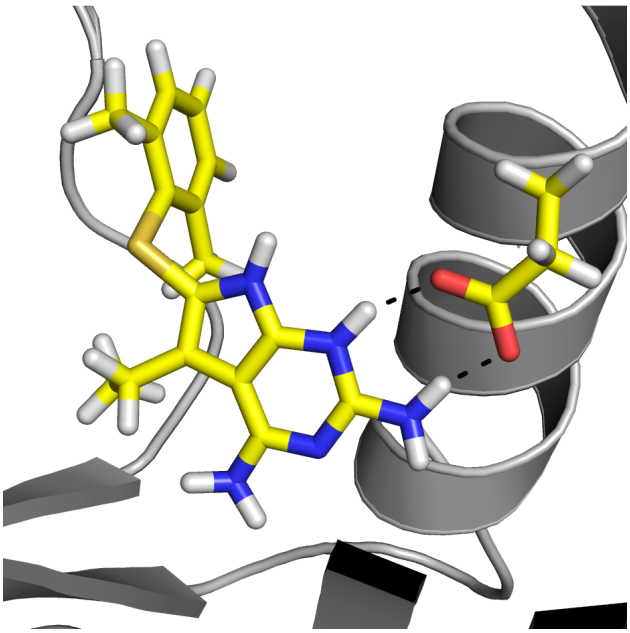


Figure 2: Hydrogen bond interactions between one of the protonated DHFR ligands and Glu30 (PDB code: 5HPB).

the Maxwell-Boltzmann distribution.

All simulations were run at 298 K using the Langevin thermostat[76] with $\tau_T = 1$ ps^{-1}. Equilibration pressure control at 1 bar was achieved with the Berendsen barostat,[77] whereas the production barostat of choice was the Parrinello-Rahman[78] barostat with $\tau_P = 1$ ps^{-1}. In all cases bonds involving hydrogen were constrained using the 4$^{\text{th}}$ order LINCS algorithm.[79] Long-range electrostatics were calculated using particle mesh Ewald[80] (PME) with real space cut-off at 1.2 nm. LJ interactions also had a cut-off at 1.2 nm with a relevant energy and pressure dispersion correction. A Verlet cut-off scheme was used for neighbour list updates every 20 integration steps.

Energy difference (ΔH) readings were taken every 10 ps and were analyzed using the Bennett acceptance ratio[81] implementation in Python.[82] Since in many cases we perturbed constrained atoms, the multistate Bennett acceptance ratio (MBAR[83]) approach was not feasible, since LINCS constraint contributions to free energy differences in GROMACS 2018.4 are extrapolated from $\frac{\partial H}{\partial \lambda}$ values and are thus not suitable for non-neighbouring λ windows.

2.3 Analysis

When analyzing results from different replicates, the errors have been assumed to be approximately normally distributed, and in these cases the sample mean and its associated standard error were used to describe the data. Normality was not assumed in all other cases, e.g. when comparing different crystal structures and equilibration times. In these cases robust statistics, such as the sample median and the interquartile range, were used.

In accordance with the non-assumption of normality, non-parametric rank-based tests were used for all comparisons, namely the Mann-Whitney U test[84] for comparing two populations and the Kruskal-Wallis test[85] for comparing multiple populations. In both tests the null hypothesis is that the mean ranks of the populations are the same and the resulting p-values indicate the probability of observing the data given that the null hypothesis is correct. Since in all cases the compared populations were sampled from practically the same distribution, meaning that the null hypothesis was satisfied, the p-values were not used as a tool for statistical inference but rather as an approximate measure of apparent sampling quality. Where applicable, correlation between populations was also measured in a rank-based fashion, using Kendall's τ.[86]

In the Supporting Information, we conduct analysis which involves 145 independent Kruskal-Wallis tests. We have reported a p-value for each test but we have not attempted to demonstrate any statistical significance, since the large number of tests increases the probability of a type I error. These values have therefore been presented purely for demonstration purposes.

3 Results and Analysis

3.1 Variance Between Structures after 100 ps and 20 ns Equilibration

DHFR The calculated $\Delta\Delta G^{\ominus}$ values across all perturbed pairs and crystal structures are shown in Figures 3 and 4 after 100 ps and an additional 20 ns equilibration, respectively. It can be seen that the "largest" (i.e. perturbing the highest number of atoms) perturbation (pair 5) has the highest total variance with 90% of the samples spanning a confidence interval (CI) of ~ 2 kcal/mol. Correspondingly, the "smallest" perturbation (pair 8) has the lowest spread of $\Delta\Delta G^{\ominus}$ values with CI spanning ~ 0.5 kcal/mol at the 90% level with all other perturbations exhibiting an approximately similar variance of ~ 1 kcal/mol at this CI. This correlation between perturbation size and total data spread is hardly surprising, since the same computational time was dedicated to perturbations of varying difficulty. More notably, inter-replicate variance is generally low with no crystal structures exhibiting consistently higher variances. Therefore, the total variance is mostly explained by the inter-structure variance, meaning that the use of simulation repeats starting from the same coordinates is not capturing the variance observed when using different but acceptable crystal structures. This is exemplified by the low p-values shown in Figure 3 obtained by the non-parametric Kruskal-Wallis test. It can be noted that 6DAV is a consistent outlier in most of the cases – an unsurprising observation which can be readily explained by the different cofactor in the crystal structure ($NADP^+$, instead of NADPH). However, most inter-crystal differences are significant at the 10% level even if we discard 6DAV – a value which is still concerning, since in principle the choice of initial crystal structure should have little to no effect on the free energy values.

These observations change drastically after 20 ns pre-equilibration at $\lambda = 0$. In this case we observe increased total variances with pairs 5 and 7 exhibiting a spread of $\Delta\Delta G^{\ominus}$ values over ~ 3 kcal/mol at 90% CI. Even the perturbations with the smallest variances span a

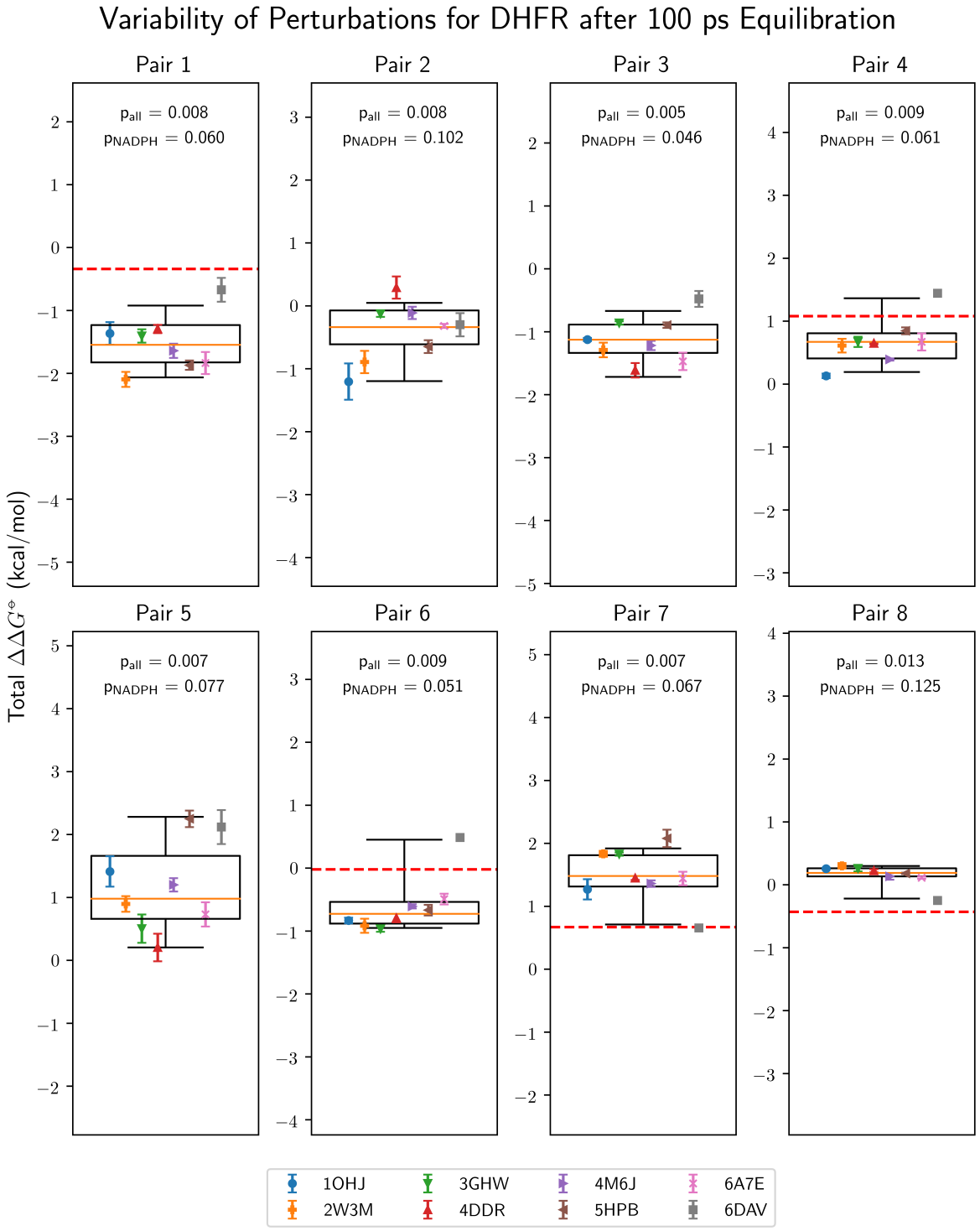


Figure 3: Box plots of the $\Delta\Delta G^{\ominus}$ values per perturbation for each of the DHFR crystal structures after 100 ps total equilibration time. Each point represents the average of three repeats and the associated error bar is its standard error of the mean. The boxes contain all values between 25$^{\text{th}}$ and 75$^{\text{th}}$ percentile and the whiskers are based on the 5$^{\text{th}}$ and 95$^{\text{th}}$ percentile. The p-values have been obtained from the Kruskal-Wallis test on all samples (p_{all}) and on all samples except for 6DAV (p_{NADPH}). The solid orange line shows the median value and the dashed red line denotes the measured experimental value,[29] if available.

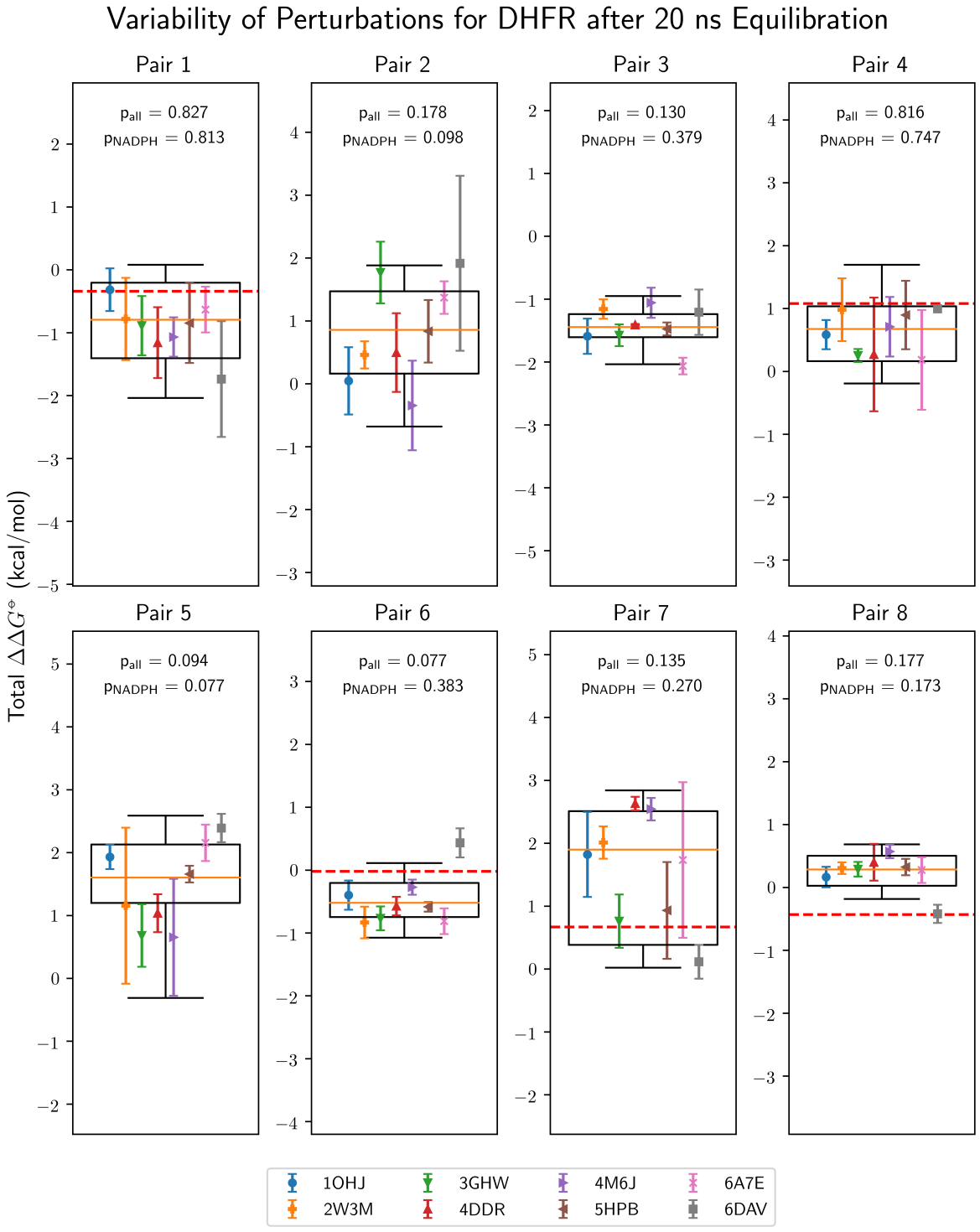


Figure 4: Box plots of the $\Delta\Delta G^{\ominus}$ values per perturbation for each of the DHFR crystal structures after 20 ns total equilibration time. Each point represents the average of three repeats and the associated error bar is its standard error of the mean. The boxes contain all values between 25$^{\text{th}}$ and 75$^{\text{th}}$ percentile and the whiskers are based on the 5$^{\text{th}}$ and 95$^{\text{th}}$ percentile. The p-values have been obtained from the Kruskal-Wallis test on all samples (p_{all}) and on all samples except for 6DAV (p_{NADPH}). The solid orange line shows the median value and the dashed red line denotes the measured experimental value,[29] if available.

range of $\sim$ 1 kcal/mol at this CI. This time much larger inter-replicate standard errors are observed with the largest one being 6DAV in pair 2 with $\sigma_{\overline{\Delta\Delta G^{\ominus}}} \approx 1.5$ kcal/mol. Although it is unsurprising that increased decorrelation between replicates results in higher variance, the magnitude of this increase after only 20 ns is remarkable. Although 6DAV is still a rather consistent outlier, it is much less so, resulting in heightened p-values, meaning that there is no significant difference between the crystal structures. Most p-values are now insignificant at the 10% level, again demonstrating the increased loss of memory of the starting crystal structure.

PTP1B The corresponding data for PTP1B can be seen in Figures 5 and 6. Similarly to DHFR, here we observe a total data spread ranging between $\sim$ 0.5 kcal/mol at the 90% CI for the simpler perturbations (pairs 7 and 8) up to $\sim$ 1.5 kcal/mol for the most difficult perturbation (pair 3). Inter-replicate variance is generally higher than for DHFR for most perturbations with no structures being consistent outliers. This is illustrated by the generally high p-values for 4 of the pairs. However, the rest of the pairs exhibit consistent significant differences at the 10% level, similar to the results obtained with DHFR. Perhaps the most curious perturbation is pair 9, which has a low p-value and a very low inter-replicate variance, exhibiting a total spread of values of over $\sim$ 1 kcal/mol. This behaviour would not have been expected if we had only used a single crystal structure which exhibits apparent convergence.

The corresponding results after a longer equilibration time are similar to those obtained for DHFR. It can be seen in Figure 6 that the total spread of $\Delta\Delta G^{\ominus}$ values is generally much larger. The most remarkable example of this is pair 5 with a total value range of more than 5 kcal/mol at 90% CI. Only the last three pairs exhibit spread of less than 1 kcal/mol, whereas all perturbations involving the sulfonamide derivative have an uncertainty of $\sim$ 3 kcal/mol at 90% CI. In all cases the inter-replicate variance is also markedly higher than the shorter runs with the highest $\sigma_{\overline{\Delta\Delta G^{\ominus}}} > 1.5$ kcal/mol (1NWE, pair 5). Moreover, the new p-values are also on average higher in comparison, once again showing that longer decorrelation times

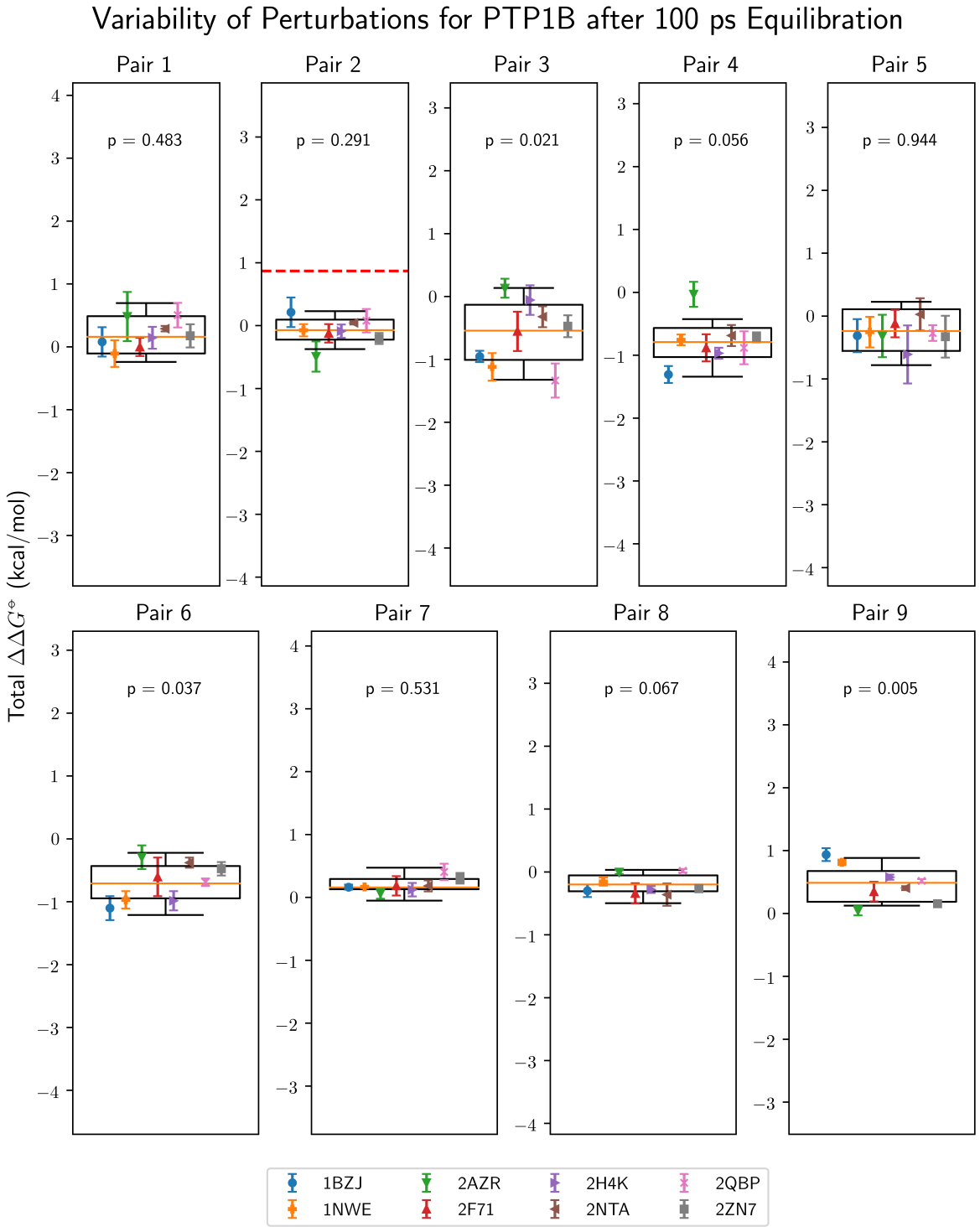


Figure 5: Box plots of the $\Delta\Delta G^{\circ}$ values per perturbation for each of the PTP1B crystal structures after 100 ps total equilibration time. Each point represents the average of three repeats and the associated error bar is its standard error of the mean. The boxes contain all values between 25$^{\text{th}}$ and 75$^{\text{th}}$ percentile and the whiskers are based on the 5$^{\text{th}}$ and 95$^{\text{th}}$ percentile. The p-values have been obtained from the Kruskal-Wallis test on all samples. The solid orange line shows the median value and the dashed red line denotes the measured experimental value,[51] if available.

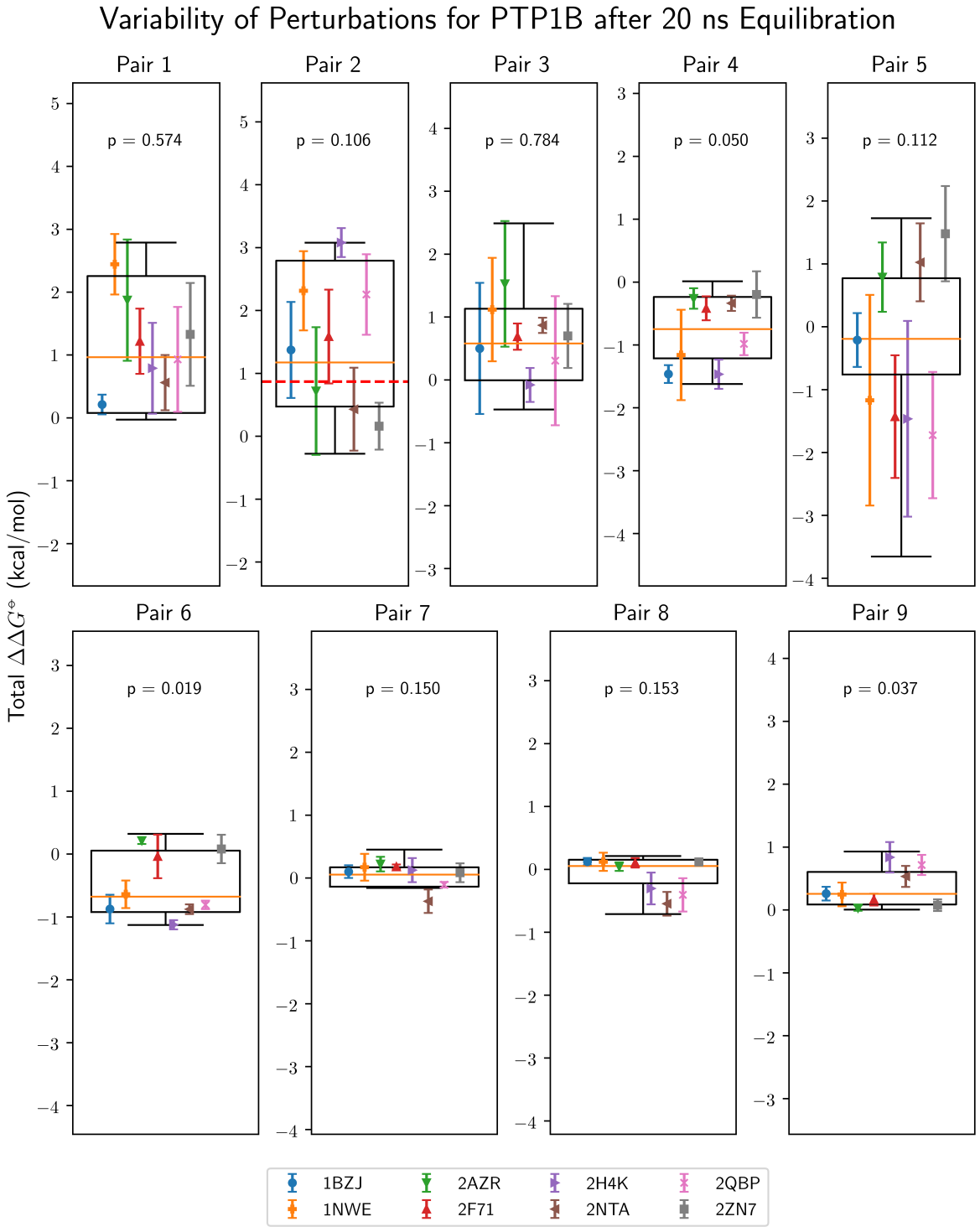


Figure 6: Box plots of the $\Delta\Delta G^{\circ}$ values per perturbation for each of the PTP1B crystal structures after 20 ns total equilibration time. Each point represents the average of three repeats and the associated error bar is its standard error of the mean. The boxes contain all values between 25$^{\text{th}}$ and 75$^{\text{th}}$ percentile and the whiskers are based on the 5$^{\text{th}}$ and 95$^{\text{th}}$ percentile. The p-values have been obtained from the Kruskal-Wallis test on all samples. The solid orange line shows the median value and the dashed red line denotes the measured experimental value,[51] if available.

result in reduced distinguishability between different crystal structures. Nevertheless, pairs 4, 6 and 9 exhibit significant differences at the 10% level, implying that the initial crystal structures still influence the obtained free energies. These observations futher demonstrate that any apparent convergence at shorter timescales is usually deceiving.

FXa The results for FXa are shown in Figures 7 and 8. In this case we generally observe smaller inter-structure variances than the previous systems with the largest spread being ~ 1 kcal/mol for the largest perturbation (pair 3) at 90% CI. In some cases this spread is less than 0.2 kcal/mol, indicating good apparent agreement between initial crystal structures. However, inter-replicate variance is generally even lower, resulting in all but two perturbation pairs being significantly different at the 10% level. For example, in pair 3 one can observe a maximum difference of ~ 1 kcal/mol for two of the crystal structures with little apparent variance, once again highlighting the impact of the choice of initial crystal structure at shorter timescales.

Similarly to previous data, FXa exhibits larger inter-structure variance across all perturbations after prolonged equilibration ranging from ~ 0.5 to ~ 1 kcal/mol. Especially curious is pair 9, which shows dramatic relative increase in variance compared to the short-equilibration results. However, the absolute spread in free energy values is still unremarkable in light of the previous test cases. In all but two perturbations, the differences between initial crystal structures are insignificant at the 10% level, once again demonstrating decreasing dependency of $\Delta\Delta G^{\ominus}$ on the choice of initial crystal structure over time, as expected.

3.2 Comparison Between $\Delta\Delta G^{\ominus}$ after 100 ps and 20 ns Equilibration

In addition to the previous analysis, we can compare the distribution of $\Delta\Delta G^{\ominus}$ values across all crystal structures and replicates after 100 ps and 20 ns equilibration. Comparison between median $\Delta\Delta G^{\ominus}$ for each pair and system is shown in Figure 9 with associated

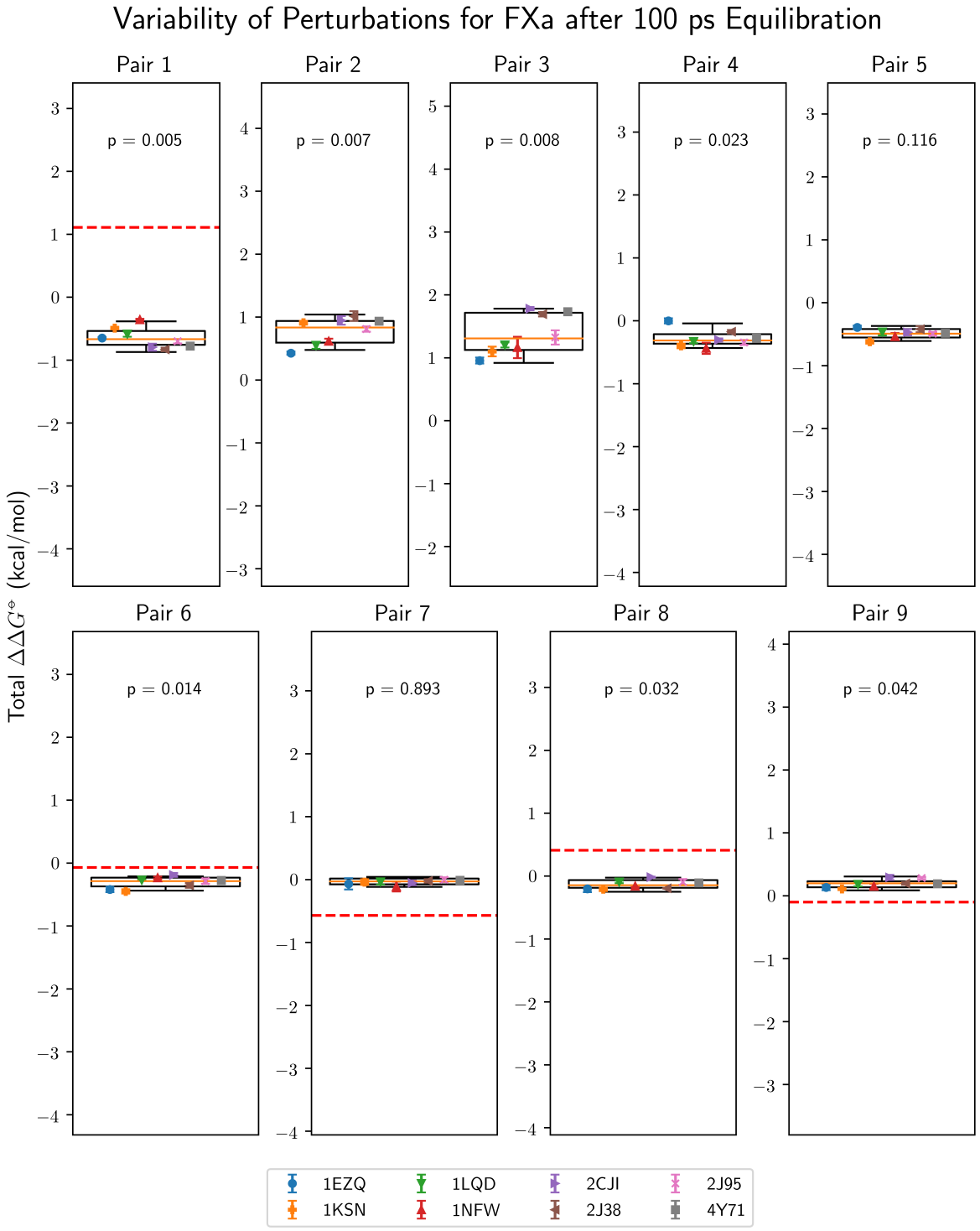


Figure 7: Box plots of the $\Delta\Delta G^{\ominus}$ values per perturbation for each of the FXa crystal structures after 100 ps total equilibration time. Each point represents the average of three repeats and the associated error bar is its standard error of the mean. The boxes contain all values between 25$^{\text{th}}$ and 75$^{\text{th}}$ percentile and the whiskers are based on the 5$^{\text{th}}$ and 95$^{\text{th}}$ percentile. The p-values have been obtained from the Kruskal-Wallis test on all samples. The solid orange line shows the median value and the dashed red line denotes the measured experimental value,[55] if available.

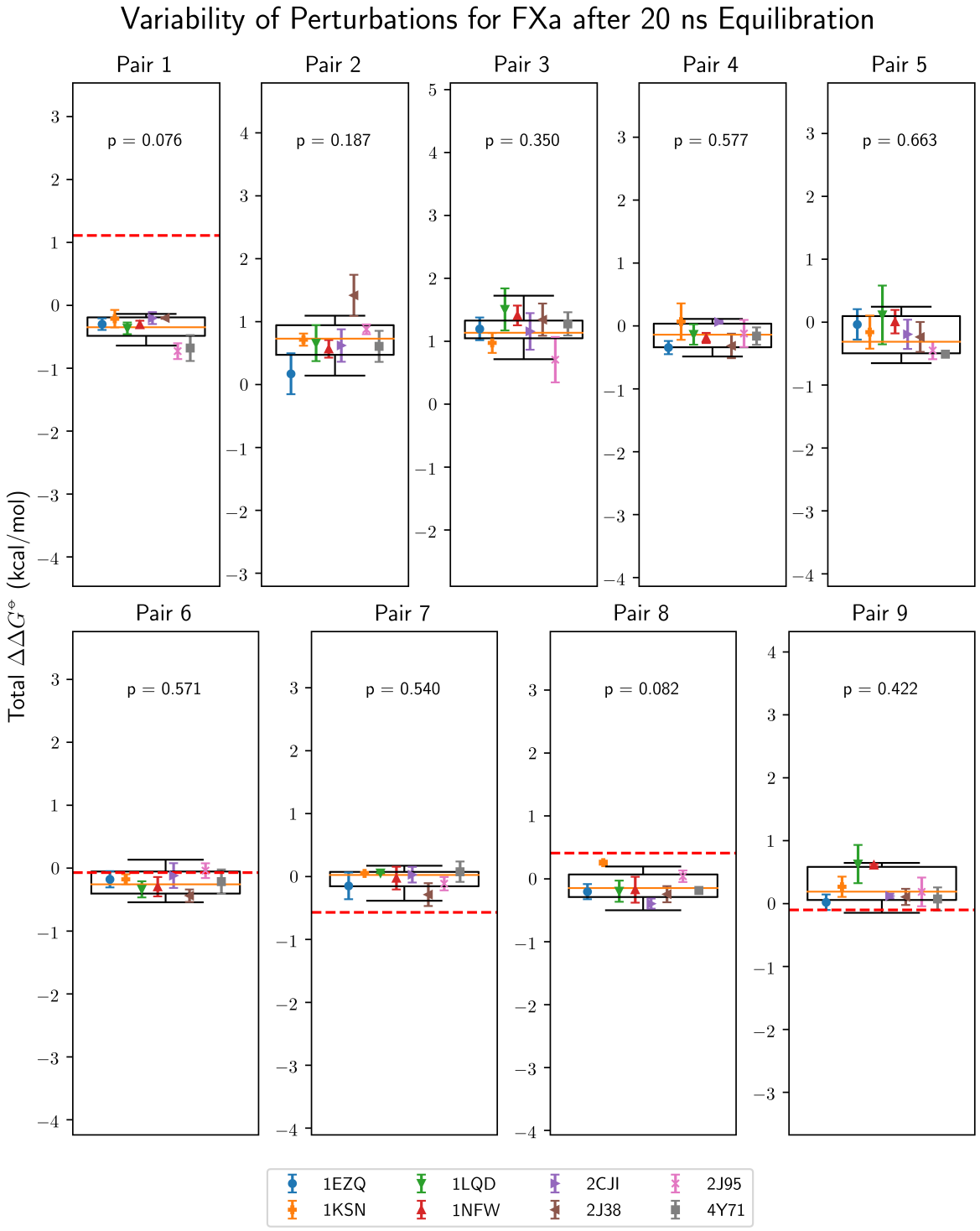
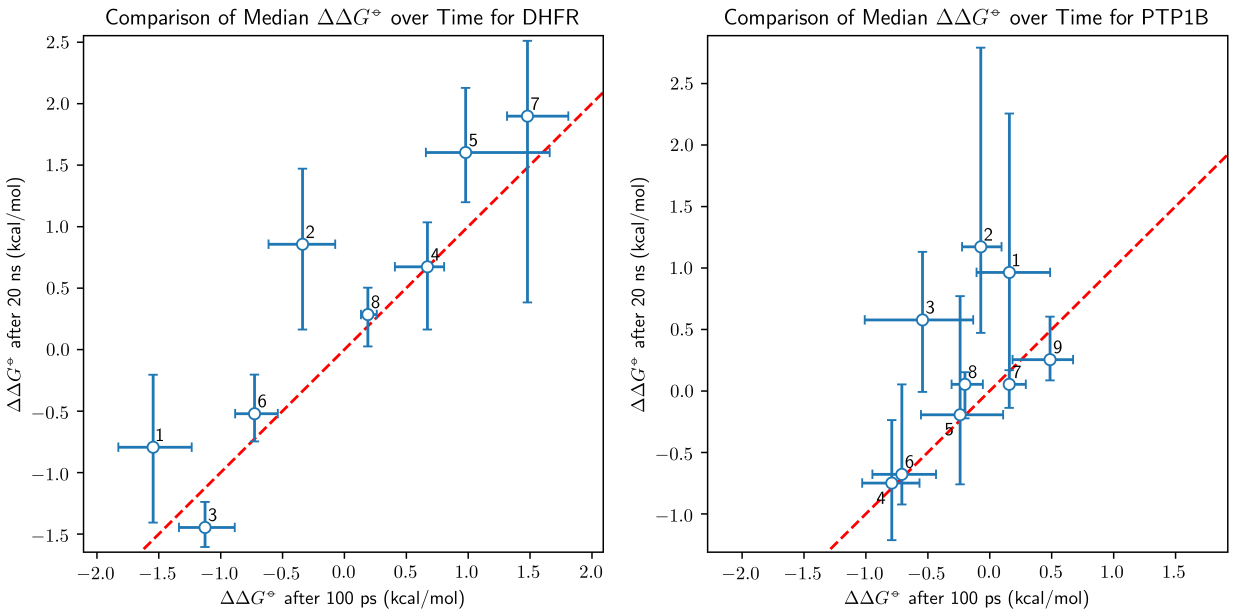


Figure 8: Box plots of the $\Delta\Delta G^{\ominus}$ values per perturbation for each of the FXa crystal structures after 20 ns total equilibration time. Each point represents the average of three repeats and the associated error bar is its standard error of the mean. The boxes contain all values between 25$^{\text{th}}$ and 75$^{\text{th}}$ percentile and the whiskers are based on the 5$^{\text{th}}$ and 95$^{\text{th}}$ percentile. The p-values have been obtained from the Kruskal-Wallis test on all samples. The solid orange line shows the median value and the dashed red line denotes the measured experimental value,[55] if available.

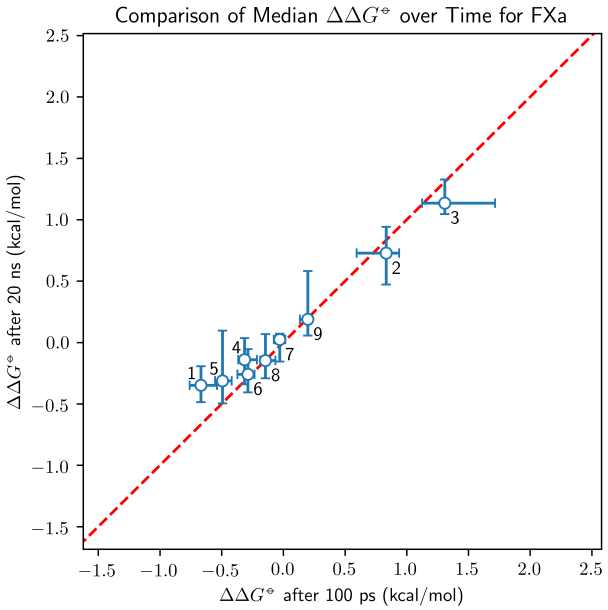
Mann-Whitney U test[84] p-values in the Supporting Information (Table S4). In some cases we see remarkable median differences in the calculated free energies of approximately $1-1.5$ kcal/mol (DHFR: pair 2; PTP1B: pairs 1, 2 and 3), whereas most other values, including all of the FXa perturbations, appear to approximately agree by visual inspection. However, the Mann-Whitney U test indicates significant differences at 2% CI for: DHFR pairs 1, 2 and 3; PTP1B pairs 1, 2, 3, 7 and 8; FXa pairs 1 and 5, which constitute more than a third of all perturbations. This indicates that even after comparing across protein crystal structures and repeats we observe significant time-dependent sampling changes. Nevertheless, it has to be noted that these differences could to some extent arise from the sampling bias introduced by prolonged equilibration at only a single λ value and one should ideally compare datasets where all λ values have been independently equilibrated for 20 ns. In our study, this was not feasible due to computational resource limitations.

3.3 Cycle Closure Errors

Since the Gibbs free energy G is a state function, any combination of perturbations which returns to the initial state must yield a net free energy change of zero. Any deviation from this value indicates insufficient sampling and lack of convergence. It can be seen (Table 1) that in most cases cycle closure errors indicate apparent convergence (less than 1 kcal/mol) after a short equilibration both on a crystal-by-crystal basis and on average, with the notable exception of all cycles involving PTP1B and the sulfonamide ligand derivative (cycles A, B and C). However, this apparent convergence is not observed after longer equilibration with some cycle closure errors surpassing the 2 kcal/mol barrier for some crystal structures. Nevertheless, all average cycle closures are within 1 kcal/mol with the exception of DHFR, cycle B, which has a magnitude of 1.67 kcal/mol, despite exhibiting apparent convergence at shorter equilibration times. This is a striking observation, since one would expect that a net equilibration time of ~ 0.5 µs and a total sampling time of ~ 4 µs per perturbation to exhibit unconditional convergence, especially for these rather straightforward perturbations. These



(a) (b)



(c)

Figure 9: Comparison of median $\Delta\Delta G^{\circ}$ values across all initial crystal structures and replicates after short (100 ps) and long (20 ns) equilibration for DHFR (Figure 9a), PTP1B (Figure 9b) and FXa (Figure 9c). All error bars indicate 25%–75% CI. The dashed red line represents the line $y = x$.

results show that any apparent convergence at shorter timescales can be deceiving even for simple systems and low cycle closure errors do not necessarily imply sufficient sampling.

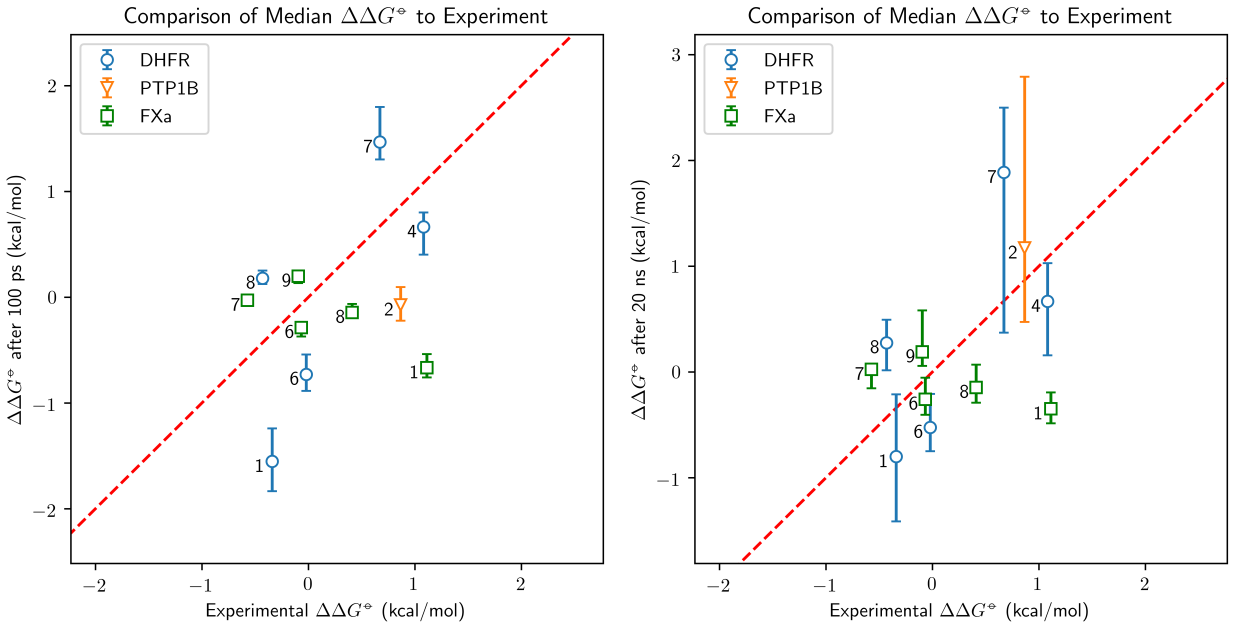
Table 1: Absolute cycle closure errors for all systems after 100 ps and 20 ns equilibration. The cycles have been calculated per structure as the average of three replicates and denoted according to Figure 1. The three columns represent the cycle closure errors from the best- and worst-performing crystal structures, as well as the average cycle closure errors between all structures.

		Cycle Closure Errors (kcal/mol)					
		Minimum		Maximum		Average	
System	Cycle	100 ps	20 ns	100 ps	20 ns	100 ps	20 ns
DHFR	A	0.03	0.07	0.97	2.44	0.02	0.30
	B	0.08	0.40	0.76	3.35	0.08	1.67
	C	0.01	0.03	0.38	1.39	0.06	0.24
PTP1B	A	0.72	0.05	1.31	1.37	1.06	0.47
	B	0.58	0.08	2.11	2.50	1.05	0.80
	C	0.01	0.00	1.02	2.03	0.18	0.28
	D	0.03	0.02	0.60	0.84	0.20	0.11
FXa	A	0.02	0.01	0.30	0.90	0.06	0.12
	B	0.01	0.21	0.17	0.53	0.03	0.05
	C	0.01	0.01	0.19	0.84	0.03	0.26

3.4 Comparison to Experiment

While not the focus of this study, which is concerned with reproducibility and precision, rather than accuracy, it is nevertheless informative to compare the above results to experimental $\Delta\Delta G^{\circ}$ values. Here we only compare direct perturbations against experiment, as opposed to thermodynamic chains. It is shown in Figure 10 that the extensively equilibrated median binding free energies generally move slightly closer to experimental values over time. The relative ranking, represented by Kendall's τ[86] changes insignificantly from 0.09 to 0.13, indicating very weak correlation to experimental data. The mean absolute deviation (MAD) also improves weakly with more equilibration from 0.73 to 0.61 kcal/mol. Both of these metrics are influenced by the low experimental free energy magnitudes ($\sim$ 1 kcal/mol), meaning

that the MAD is more likely to appear favourable and the relative ranking is dominated by noise. Owing to the size of the dataset, the low magnitude of the experimental free energy values and the high variability between different replicates, we conclude that the achieved improvement in comparison to experiment over time is not substantial for our test cases.



(a) (b)

Figure 10: Comparison of median $\Delta\Delta G^{\circ}$ values across all initial crystal structures and replicates for some of the denoted pairs after short (100 ps, Figure 10a) and long (20 ns, Figure 10b) against experiment.[29,51,55] The associated error bars indicate 25%–75% CI and the dashed red line represents the line $y = x$.

3.5 The Origin of Long-Timescale Variance

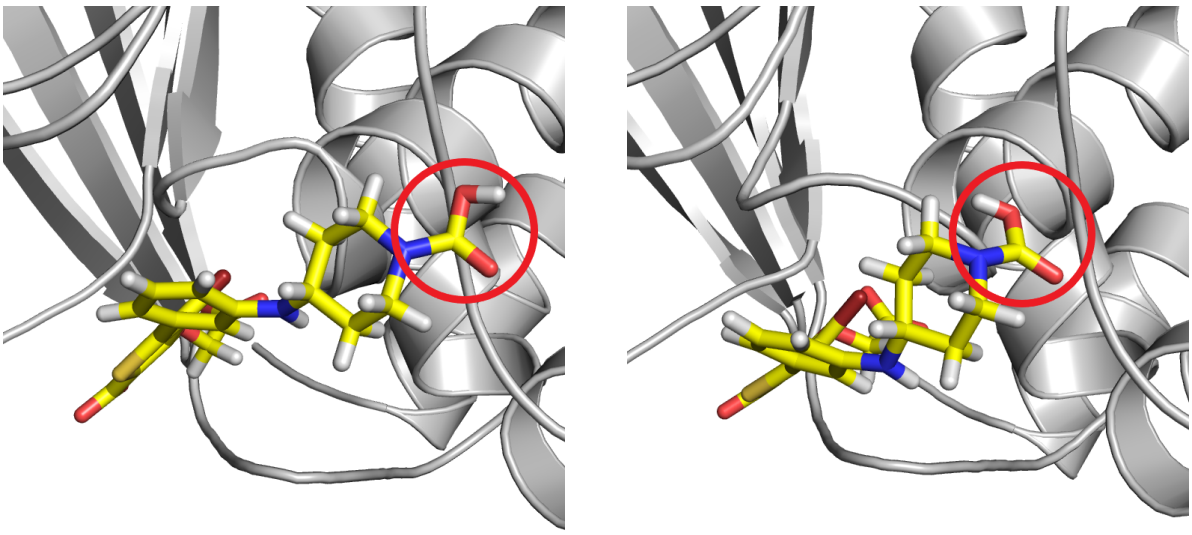
Owing to the complex nature of biomolecular systems, it is difficult to narrow down the reason for the observed increase in inter-replicate variance with simulation time. One obvious way to compare different replicates is to quantify the changes which occur during the 20 ns equilibration based on the final trajectory frame. The most apparent differences by visual inspection indicate the presence of various rare events with transition times larger than several nanoseconds, such as rotations of ligand torsions with high kinetic barriers. The

most striking example is the rotation of the acidic hydrogen in PTP1B, pair 5 (Figure 11). Comparison of free energy calculations starting from the two different rotamers reveal a median difference of more than 4 kcal/mol, indicating that this rotation is likely the primary reason for the extraordinary variance, observed in Figure 6. Another conspicuous example of rare events determining the outcome of a free energy calculation is the sulfonamide bond rotation in the first three PTP1B perturbations (Figure 12), with each rotamer exhibiting an average of approximately 1–2 kcal/mol difference to the other rotamers. Detailed analysis of these and all other rotamers can be found in the Supporting Information (Figures S1 to S4), revealing the prevalence of this trend in many of the perturbations involving DHFR and PTP1B. This analysis also shows that such ligand flexibility is observed to a much lesser extent for FXa, thereby explaining the comparatively low free energy variance even after extended equilibration.

Naturally, it is expected that the protein backbone also has an impact on the increased free energy variance over time. However, analysing such a high-dimensional dataset requires an immense amount of data points in the form of $\Delta\Delta G^{\ominus}$ values. With only 24 $\Delta\Delta G^{\ominus}$ values per perturbation, establishing a statistically significant connection between protein internal degrees of freedom and calculated free energies is not feasible and we attribute most of the long-timescale variability to slow ligand motions – a conclusion, which is supported by all of the data presented above.

4 Discussion

There are several important lessons to be learned from the above analysis. Most strikingly, we observed that at short timescales different protein crystal structures can disagree significantly over the free energy change and these differences can be more than 1 kcal/mol in magnitude. Such short-timescale simulations are commonly used in practice, most notably in commercial implementations,[32,87] making these results highly relevant to state-of-the-art applications

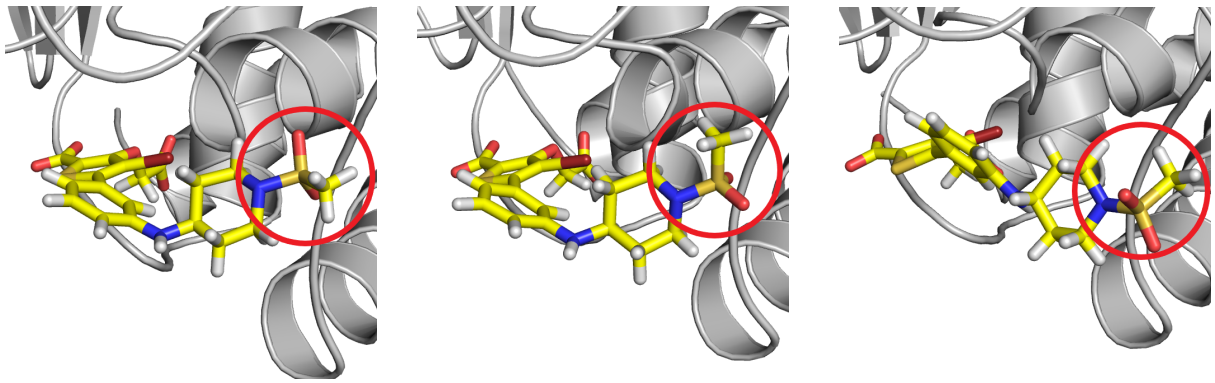


(a) (b)

Figure 11: Acidic hydrogen rotation in pair 5 observed in extended PTP1B simulations. Images generated from the final trajectory frame of the extended equilibration for 1BZJ (Figure 11a) and 1NWE (Figure 11b).

of alchemical methods. More worryingly, most of these results appear well-converged, as evidenced by the low inter-replicate variance and the satisfactory cycle closure errors. The issue of using a single crystal structure is now apparent: this choice can covertly affect the relative ranking of compounds, even when the free energy changes appear to be too large for this to be likely to occur.

As we saw above, these inter-structure differences are largely reduced after a prolonged equilibration time and the inter-triplicate differences become more representative of their true uncorrelated values. However, even after $\sim$ 24 ns of dynamics we can often distinguish between the different initial structures selected, showing that this initial choice affects our results in the long run. Nevertheless, these results show that the proposition that multiple short simulations are preferable to a single long simulation does not necessarily capture the full nature of the problem: replicates are necessary but insufficient for convergence, while many important long-timescale motions are practically inaccessible at shorter timescales, regardless of the number of replicates. Therefore, one needs both multiple replicates for



(a) (b) (c)

Figure 12: Sulfonamide rotation in pair 3 observed in extended PTP1B simulations. Images generated from the final trajectory frame of the extended equilibration for 1BZJ (Figure 12a), 2AZR (Figure 12b) and 2H4K (Figure 12c).

statistical confidence and longer simulations for physical validity – a requirement which is rarely practically feasible with current computational capabilities.

One obvious way to practically circumvent this problem is to run short simulations over more than one crystal structure. While the average of the resulting free energy values would likely be biased, at least the researcher would be aware of the minimal uncertainty in their results. However, this approach would result in reduced quality of ligand sampling, since pure molecular dynamics is not good at exploring multiple binding modes at short timescales.[88] Alternatively, one could run repeats over several binding poses determined by e.g. docking, but the same problem of determining the correct weights of each binding pose persists, resulting in biased free energy differences.

Another possibly more preferable working alternative is running longer simulations on one crystal structure with enhanced sampling of the ligand degrees of freedom. The results from longer timescales are typically more sensitive to ligand conformational changes and the initial crystal structure becomes less influential on the free energy changes over time. A protocol combining AFE and replica exchange with solute scaling (REST2[89]) has long been used in commercial implementations, such as FEP+[32] and has recently been used over longer timescales.[12,90,91] Although it is expected that this approach would lead to a much higher

variance due to the number of ligand degrees of freedom and decreased phase space overlap between neighbouring states due to the Hamiltonian rescaling, this variance would be more representative of the true result and this approach would be much less likely to exhibit false convergence. In all cases it is highly recommended to run at least triplicate simulations.

We also observed that larger perturbations result in much more variable free energy estimates, a largely expected result. However, even the simplest of perturbations should be treated with care. More specifically, cycle closure errors can indicate false convergence and should therefore only be used to demonstrate insufficient sampling. Indeed, we saw that extensive sampling usually results in higher and more realistic cycle closure errors, meaning that this criterion is necessary but not sufficient for convergence.

Since it was unclear from our data whether prolonged equilibration affects comparison to experiment significantly, one might argue that better sampling might not be necessarily cost-efficient for applications. While this is possible considering the accuracy provided by current force fields, it has to be remembered that all computational time saved from less sampling, results in reduced physical and statistical confidence. Therefore, our advice is at the very least to use timescales and enhanced sampling techniques providing sufficient ligand conformational sampling whenever possible, so that the binding conformation is not completely dependent on the ligand alignment method and/or the researcher's intuition.

5 Conclusion

We have shown the influence of initial crystal structure and extended equilibration time on the binding free energy values of three different systems: DHFR, PTP1B and FXa. Our observations indicate that at short timescales, initial crystal structure differences consistently result in statistically, although not necessarily numerically, significant changes in $\Delta\Delta G^{\ominus}$, sometimes reaching differences of over 1 kcal/mol. Furthermore, large perturbations result in higher sensitivity to the initial structure at short timescales.

At longer timescales, we observe increased inter-replicate variance in $\Delta\Delta G^{\ominus}$, which makes the results from different initial crystal structures appear more similar. In many cases, a significant contributor to this variance is the rare changes in ligand conformation which become more common at these timescales. The extent to which the protein degrees of freedom impact these results is not clear and is to be investigated in future work. Sometimes this prolonged sampling can significantly change the expected free energy value. However, we have inconclusive evidence as to whether these changes improve comparison to experiment. In addition, we demonstrated that thermodynamic cycle closure values can often indicate false convergence at short timescales, meaning that long-timescale enhanced sampling is needed even for simple perturbations.

Finally, we have justified the use of long-timescale dynamics and the use of enhanced sampling in AFE calculations as well as performing multiple repeats with the same initial configurations, meaning that an optimal protocol needs to find the balance between the number of repeats, simulation length, and the number of λ windows in the general case.

Acknowledgement

All of the presented data were generated thanks to the generous computational time on the Iridis 5 cluster provided by the University of Southampton. This study has been funded by AstraZeneca, GSK and Syngenta. We are also supported by the EPSRC Centre for Doctoral Training, Theory and Modelling in Chemical Sciences, under Grant EP/L015722/1.

Supporting Information Available

The following files are available free of charge.

- Additional tables and plots: crystal structure information, p-values and rotamer clustering analysis (PDF)

- Python scripts used to perform the study using ProtoCaller (ZIP)

References

(1) Sherborne, B.; Shanmugasundaram, V.; Cheng, A. C.; Christ, C. D.; DesJarlais, R. L.; Duca, J. S.; Lewis, R. A.; Loughney, D. A.; Manas, E. S.; McGaughey, G. B.; Peishoff, C. E.; van Vlijmen, H. Collaborating to improve the use of free-energy and other quantitative methods in drug discovery. *J. Comput. Aided Mol. Des.* **2016**, *30*, 1139–1141.

(2) Cournia, Z.; Allen, B.; Sherman, W. Relative binding free energy calculations in drug discovery: recent advances and practical considerations. *J. Chem. Inf. Model.* **2017**, *57*, 2911–2937.

(3) Williams-Noonan, B. J.; Yuriev, E.; Chalmers, D. K. Free energy methods in drug design: prospects of "alchemical perturbation" in medicinal chemistry: miniperspective. *J. Med. Chem.* **2018**, *61*, 638–649.

(4) Rocklin, G. J.; Mobley, D. L.; Dill, K. A. Calculating the sensitivity and robustness of binding free energy calculations to force field parameters. *J. Chem. Theory Comput.* **2013**, *9*, 3072–3083.

(5) Manzoni, F.; Ryde, U. Assessing the stability of free-energy perturbation calculations by performing variations in the method. *J. Comput. Aided Mol. Des.* **2018**, *32*, 529–536.

(6) Vassetti, D.; Pagliai, M.; Procacci, P. Assessment of GAFF2 and OPLS-AA general force fields in combination with the water models TIP3P, SPCE and OPC3 for the solvation free energy of drug-like organic molecules. *J. Chem. Theory Comput.* **2019**, *15*, 1983–1995.

(7) Pérez-Benito, L.; Casajuana Martin, N.; Jiménez-Rosés, M.; Van Vlijmen, H.; Tresadern, G. Predicting activity cliffs with free-energy perturbation. *J. Chem. Theory Comput.* **2019**, *15*, 1884–1895.

(8) Wan, S.; Tresadern, G.; Pérez-Benito, L.; van Vlijmen, H.; Coveney, P. V. Accuracy and precision of alchemical relative free energy predictions with and without replica-exchange. *Advanced Theory and Simulations* **2020**, *3*, 190–195.

(9) Bhati, A. P.; Wan, S.; Hu, Y.; Sherborne, B.; Coveney, P. V. Uncertainty quantification in alchemical free energy methods. *J. Chem. Theory Comput.* **2018**, *14*, 2867–2880.

(10) Rizzi, A.; Jensen, T.; Slochower, D. R.; Aldeghi, M.; Gapsys, V.; Ntekoumes, D.; Bosisio, S.; Papadourakis, M.; Henriksen, N. M.; De Groot, B. L.; Cournia, Z.; Dickson, A.; Michel, J.; Gilson, M. K.; Shirts, M. R.; Mobley, D. L.; Chodera, J. D. The SAMPL6 SAMPLing challenge: assessing the reliability and efficiency of binding free energy calculations. *J. Comput. Aided Mol. Des.* **2020**, 1–33.

(11) Hahn, D. F.; König, G.; Hünenberger, P. H. Overcoming orthogonal barriers in alchemical free energy calculations: on the relative merits of λ-variations, λ-extrapolations, and biasing. *J. Chem. Theory Comput.* **2020**, *16*, 1630–1645.

(12) Keränen, H.; Pérez-Benito, L.; Ciordia, M.; Delgado, F.; Steinbrecher, T. B.; Oehlrich, D.; van Vlijmen, H. W. T.; Trabanco, A. A.; Tresadern, G. Acylguanidine beta secretase 1 inhibitors: a combined experimental and free energy perturbation study. *J. Chem. Theory Comput.* **2017**, *13*, 1439–1453.

(13) Fratev, F.; Sirimulla, S. An improved free energy perturbation FEP+ sampling protocol for flexible ligand-binding domains. *Sci. Rep.* **2019**, *9*, 1–13.

(14) Michel, J.; Verdonk, M. L.; Essex, J. W. Protein-ligand binding affinity predictions by implicit solvent simulations: a tool for lead optimization? *J. Med. Chem.* **2006**, *49*, 7427–7439.

(15) Pan, A. C.; Xu, H.; Palpant, T.; Shaw, D. E. Quantitative characterization of the binding and unbinding of millimolar drug fragments with molecular dynamics simulations. *J. Chem. Theory Comput.* **2017**, *13*, 3372–3377.

(16) Shirts, M. R.; Pande, V. S. Comparison of efficiency and bias of free energies computed by exponential averaging, the Bennett acceptance ratio, and thermodynamic integration. *J. Chem. Phys.* **2005**, *122*, 144107.

(17) Shirts, M. R.; Klein, C.; Swails, J. M.; Yin, J.; Gilson, M. K.; Mobley, D. L.; Case, D. A.; Zhong, E. D. Lessons learned from comparing molecular dynamics engines on the SAMPL5 dataset. *J. Comput. Aided Mol. Des.* **2017**, *31*, 147–161.

(18) Loeffler, H. H.; Bosisio, S.; Duarte Ramos Matos, G.; Suh, D.; Roux, B.; Mobley, D. L.; Michel, J. Reproducibility of free energy calculations across different molecular simulation software packages. *J. Chem. Theory Comput.* **2018**, *14*, 5567–5582.

(19) Michel, J.; Essex, J. W. Prediction of protein–ligand binding affinity by free energy simulations: assumptions, pitfalls and expectations. *J. Comput. Aided Mol. Des.* **2010**, *24*, 639–658.

(20) Chen, W.; Deng, Y.; Russell, E.; Wu, Y.; Abel, R.; Wang, L. Accurate calculation of relative binding free energies between ligands with different net charges. *J. Chem. Theory Comput.* **2018**, *14*, 6346–6358.

(21) de Oliveira, C.; Yu, H. S.; Chen, W.; Abel, R.; Wang, L. Rigorous free energy perturbation approach to estimating relative binding affinities between ligands with multiple protonation and tautomeric states. *J. Chem. Theory Comput.* **2018**, *15*, 424–435.

(22) Luccarelli, J.; Michel, J.; Tirado-Rives, J.; Jorgensen, W. L. Effects of water placement on predictions of binding affinities for p38α MAP kinase inhibitors. *J. Chem. Theory Comput.* **2010**, *6*, 3850–3856.

(23) Bodnarchuk, M. S.; Viner, R.; Michel, J.; Essex, J. W. Strategies to calculate water binding free energies in protein–ligand complexes. *J. Chem. Inf. Model.* **2014**, *54*, 1623–1633.

(24) Bruce Macdonald, H. E.; Cave-Ayland, C.; Ross, G. A.; Essex, J. W. Ligand binding free energies with adaptive water networks: two-dimensional grand canonical alchemical perturbations. *J. Chem. Theory Comput.* **2018**, *14*, 6586–6597.

(25) Wahl, J.; Smieško, M. Assessing the predictive power of relative binding free energy calculations for test cases involving displacement of binding site water molecules. *J. Chem. Inf. Model.* **2019**, *59*, 754–765.

(26) Granadino-Roldán, J. M.; Mey, A. S.; Pérez González, J. J.; Bosisio, S.; Rubio-Martinez, J.; Michel, J. Effect of set up protocols on the accuracy of alchemical free energy calculation over a set of ACK1 inhibitors. *PLoS One* **2019**, *14*, e0213217.

(27) Cappel, D.; Jerome, S.; Hessler, G.; Matter, H. Impact of Different Automated Binding Pose Generation Approaches on Relative Binding Free Energy Simulations. *J. Chem. Inf. Model.* **2020**, *60*, 1432–1444.

(28) Lim, N. M.; Wang, L.; Abel, R.; Mobley, D. L. Sensitivity in binding free energies due to protein reorganization. *J. Chem. Theory Comput.* **2016**, *12*, 4620–4631.

(29) Tosso, R. D.; Andujar, S. A.; Gutierrez, L.; Angelina, E.; Rodríguez, R.; Nogueras, M.; Baldoni, H.; Suvire, F. D.; Cobo, J.; Enriz, R. D. Molecular modeling study of dihydrofolate reductase inhibitors. Molecular dynamics simulations, quantum mechanical calculations, and experimental corroboration. *J. Chem. Inf. Model.* **2013**, *53*, 2018–2032.

(30) Genheden, S.; Nilsson, I.; Ryde, U. Binding affinities of factor Xa inhibitors estimated by thermodynamic integration and MM/GBSA. *J. Chem. Inf. Model.* **2011**, *51*, 947–958.

(31) Genheden, S. Are homology models sufficiently good for free-energy simulations? *J. Chem. Inf. Model.* **2012**, *52*, 3013–3021.

(32) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, L.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.

(33) Knapp, B.; Ospina, L.; Deane, C. M. Avoiding false positive conclusions in molecular simulation: the importance of replicas. *J. Chem. Theory Comput.* **2018**, *14*, 6127–6138.

(34) Chen, J.; Wang, J.; Yin, B.; Pang, L.; Wang, W.; Zhu, W. Molecular mechanism of binding selectivity of inhibitors toward BACE1 and BACE2 revealed by multiple short molecular dynamics simulations and free-energy predictions. *ACS Chem. Neurosci.* **2019**, *10*, 4303–4318.

(35) Wright, D. W.; Wan, S.; Meyer, C.; Van Vlijmen, H.; Tresadern, G.; Coveney, P. V. Application of ESMACS binding free energy protocols to diverse datasets: bromodomain-containing protein 4. *Sci. Rep.* **2019**, *9*, 1–15.

(36) Suruzhon, M.; Senapathi, T.; Bodnarchuk, M. S.; Viner, R.; Wall, I.; Barnett, C. B.; Naidoo, K. J.; Essex, J. W. ProtoCaller: robust automation of binding free energy calculations. *J. Chem. Inf. Model.* **2020**, *60*, 1917–1921.

(37) Cody, V.; Galitsky, N.; Luft, J. R.; Pangborn, W.; Rosowsky, A.; Blakley, R. L. Comparison of two independent crystal structures of human dihydrofolate reductase ternary complexes reduced with nicotinamide adenine dinucleotide phosphate and the very tight-binding inhibitor PT523. *Biochemistry* **1997**, *36*, 13897–13903.

(38) Leung, A. K. W.; Ross, L. J.; Zywno-Van Ginkel, S.; Reynolds, R. C.; Seitz, L. E.; Pathak, V.; Barrow, W. W.; White, E. L.; Suling, W. J.; Piper, J. R.; Borhani, D. W. Structural basis for selective inhibition of Mycobacterium avium dihydrofolate reductase by a lipophilic antifolate. Protein Data Bank, 2009.

(39) Gangjee, A.; Li, W.; Kisliuk, R. L.; Cody, V.; Pace, J.; Piraino, J.; Makin, J. Design, synthesis, and X-ray crystal structure of classical and nonclassical 2-amino-4-oxo-5-substituted-6-ethylthieno [2, 3-d] pyrimidines as dual thymidylate synthase and dihydrofolate reductase inhibitors and as potential antitumor agents. *J. Med. Chem.* **2009**, *52*, 4892–4902.

(40) Yuthavong, Y.; Tarnchompoo, B.; Vilaivan, T.; Chitnumsub, P.; Kamchonwongpaisan, S.; Charman, S. A.; McLennan, D. N.; White, K. L.; Vivas, L.; Bongard, E.; Thongphanchang, C.; Taweechai, S.; Vanichtanankul, J.; Rattanajak, R.; Arwon, U.; Fantauzzi, P.; Yuvaniyama, J.; Charman, W.; Matthews, D. Malarial dihydrofolate reductase as a paradigm for drug development against a resistance-compromised target. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 16823–16828.

(41) Bhabha, G.; Ekiert, D. C.; Jennewein, M.; Zmasek, C. M.; Tuttle, L. M.; Kroon, G.; Dyson, H. J.; Godzik, A.; Wilson, I. A.; Wright, P. E. Divergent evolution of protein conformational dynamics in dihydrofolate reductase. *Nat. Struct. Mol. Biol.* **2013**, *20*, 1243.

(42) Cody, V.; Gangjee, A. Human dihydrofolate reductase ternary complex with a series offluorine substituted 5-methyl-6-phenythio)thieno[2,3-d]pyrimidine-2,4-diamines. Protein Data Bank, 2016.

(43) Tarnchompoo, B.; Chitnumsub, P.; Jaruwat, A.; Shaw, P. J.; Vanichtanankul, J.; Poen, S.; Rattanajak, R.; Wongsombat, C.; Tonsomboon, A.; Decharuangsilp, S.; Anukunwithaya, T.; Arwon, U.; Kamchonwongpaisan, S.; Yuthavong, Y. Hybrid in-

hibitors of malarial dihydrofolate reductase with dual binding modes that can forestall resistance. *ACS Med. Chem. Lett.* **2018**, *9*, 1235–1240.

(44) Mayclin, S. J.; Dranow, D. M.; Walpole, C.; Lorimer, D. D. Crystal Structure of Human DHFR complexed with NADP and N10-formyltetrahydrofolate. Protein Data Bank, 2018.

(45) Groves, M. R.; Yao, Z.-J.; Roller, P. P.; Burke, T. R.; Barford, D. Structural basis for inhibition of the protein tyrosine phosphatase 1B by phosphotyrosine peptide mimetics. *Biochemistry* **1998**, *37*, 17773–17783.

(46) Erlanson, D. A.; McDowell, R. S.; He, M. M.; Randal, M.; Simmons, R. L.; Kung, J.; Waight, A.; Hansen, S. K. Discovery of a new phosphotyrosine mimetic for PTP1B using breakaway tethering. *J. Am. Chem. Soc.* **2003**, *125*, 5602–5603.

(47) Moretto, A. F.; Kirincich, S. J.; Xu, W. X.; Smith, M. J.; Wan, Z.-K.; Wilson, D. P.; Follows, B. C.; Binnun, E.; Joseph-McCarthy, D.; Foreman, K.; Erbe, D. V.; Zhang, Y. L.; Tam, S. K.; Tam, S. Y.; Lee, J. Bicyclic and tricyclic thiophenes as protein tyrosine phosphatase 1B inhibitors. *Bioorg. Med. Chem.* **2006**, *14*, 2162–2177.

(48) Klopfenstein, S. R.; Evdokimov, A. G.; Colson, A.-O.; Fairweather, N. T.; Neuman, J. J.; Maier, M. B.; Gray, J. L.; Gerwe, G. S.; Stake, G. E.; Howard, B. W.; Farmer, J. A.; Pokross, M. E.; Downs, T. R.; Kasibhatla, B.; Peters, K. G. 1, 2, 3, 4-Tetrahydroisoquinolinyl sulfamic acids as phosphatase PTP1B inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1574–1578.

(49) Wan, Z.-K.; Lee, J.; Xu, W.; Erbe, D. V.; Joseph-McCarthy, D.; Follows, B. C.; Zhang, Y.-L. Monocyclic thiophenes as protein tyrosine phosphatase 1B inhibitors: capturing interactions with Asp48. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 4941–4945.

(50) Wan, Z.-K.; Follows, B.; Kirincich, S.; Wilson, D.; Binnun, E.; Xu, W.; Joseph-McCarthy, D.; Wu, J.; Smith, M.; Zhang, Y.-L.; Tam, M.; Erbe, D.; Tam, S.; Saiah, E.;

Lee, J. Probing acid replacements of thiophene PTP1B inhibitors. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 2913–2920.

(51) Wilson, D. P.; Wan, Z.-K.; Xu, W.-X.; Kirincich, S. J.; Follows, B. C.; Joseph-McCarthy, D.; Foreman, K.; Moretto, A.; Wu, J.; Zhu, M.; Binnun, E.; Zhang, Y. L.; Tam, M.; Erbe, D. V.; Tobin, J.; Xu, X.; Leung, L.; Shilling, A.; Tam, S. Y.; Mansour, T. S.; Lee, J. Structure-based optimization of protein tyrosine phosphatase 1B inhibitors: from the active site to the second phosphotyrosine binding site. *J. Med. Chem.* **2007**, *50*, 4681–4698.

(52) Wan, Z.-K.; Lee, J.; Hotchandani, R.; Moretto, A.; Binnun, E.; Wilson, D. P.; Kirincich, S. J.; Follows, B. C.; Ipek, M.; Xu, W.; ; Joseph-McCarthy, D.; Zhang, Y. L.; Tam, M.; Erbe, D. V.; Tobin, J. F.; Li, W.; Tam, S. Y.; Mansour, T. S.; Wu, J. Structure-based optimization of protein tyrosine phosphatase 1B inhibitors: capturing interactions with arginine 24. *ChemMedChem* **2008**, *3*, 1525–1529.

(53) Maignan, S.; Guilloteau, J.-P.; Pouzieux, S.; Choi-Sledeski, Y. M.; Becker, M. R.; Klein, S. I.; Ewing, W. R.; Pauls, H. W.; Spada, A. P.; Mikol, V. Crystal structures of human factor Xa complexed with potent inhibitors. *J. Med. Chem.* **2000**, *43*, 3226–3232.

(54) Guertin, K. R.; Gardner, C. J.; Klein, S. I.; Zulli, A. L.; Czekaj, M.; Gong, Y.; Spada, A. P.; Cheney, D. L.; Maignan, S.; Guilloteau, J.-P.; Brown, K. D.; Colussi, D. J.; Chu, V.; Heran, C. L.; Morgan, S. R.; Bentley, R. G.; Dunwiddie, C. T.; Leadley, R. J.; Pauls, H. W. Optimization of the β-aminoester class of factor Xa inhibitors. Part 2: identification of FXV673 as a potent and selective inhibitor with excellent in vivo anticoagulant activity. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 1671–1674.

(55) Matter, H.; Defossa, E.; Heinelt, U.; Blohm, P.-M.; Schneider, D.; Müller, A.; Herok, S.; Schreuder, H.; Liesum, A.; Brachvogel, V.; Lönze, P.; Walser, A.; Al-Obeidi, F.; Wild-

goose, P. Design and quantitative structure- activity relationship of 3-amidinobenzyl-1 H-indole-2-carboxamides as potent, nonchiral, and selective inhibitors of blood coagulation factor Xa. *J. Med. Chem.* **2002**, *45*, 2749–2769.

(56) Maignan, S.; Guilloteau, J.-P.; Choi-Sledeski, Y. M.; Becker, M. R.; Ewing, W. R.; Pauls, H. W.; Spada, A. P.; Mikol, V. Molecular structures of human factor Xa complexed with ketopiperazine inhibitors: preference for a neutral group in the S1 pocket. *J. Med. Chem.* **2003**, *46*, 685–690.

(57) Watson, N. S.; Brown, D.; Campbell, M.; Chan, C.; Chaudry, L.; Convery, M. A.; Fenwick, R.; Hamblin, J. N.; Haslam, C.; Kelly, H. A.; King, N. P.; Kurtis, C. L.; Leach, A. R.; Manchee, G. R.; Mason, A. M.; Mitchell, C.; Patel, C.; Patel, V. K.; Senger, S.; Shah, G. P.; Weston, H. E.; Whitworth, C.; Young, R. J. Design and synthesis of orally active pyrrolidin-2-one-based factor Xa inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 3784–3788.

(58) Senger, S.; Convery, M. A.; Chan, C.; Watson, N. S. Arylsulfonamides: a study of the relationship between activity and conformational preferences for a series of factor Xa inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5731–5735.

(59) Chan, C.; Borthwick, A. D.; Brown, D.; Burns-Kurtis, C. L.; Campbell, M.; Chaudry, L.; Chung, C.-w.; Convery, M. A.; Hamblin, J. N.; Johnstone, L.; Kelly, H. A.; Kleanthous, S.; Patikis, A.; Patel, C.; Pateman, A. J.; Senger, S.; Shah, G. P.; Toomey, J. R.; Watson, N. S.; Weston, H. E.; Whitworth, C.; Young, R. J.; Zhou, P. Factor Xa inhibitors: S1 binding interactions of a series of N-{(3S)-1-[(1S)-1-methyl-2-morpholin-4-yl-2-oxoethyl]-2-oxopyrrolidin-3-yl} sulfonamides. *J. Med. Chem.* **2007**, *50*, 1546–1557.

(60) Convery, M. A.; Young, R. J.; Senger, S.; Hamblin, J. N.; Chan, C.; Toomey, J. R.; Watson, N. S. Factor Xa complex with GTC000398. Protein Data Bank, 2015.

(61) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(62) Webb, B.; Sali, A. Protein structure modeling with MODELLER. In *Functional Genomics*; Springer, 2017; pp 39–54.

(63) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32*, W665–W667.

(64) Schrödinger, LLC, The PyMOL Molecular Graphics System, Version 1.8. 2015.

(65) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: an open chemical toolbox. *J. Cheminformatics* **2011**, *3*, 33.

(66) Forli, S. Charting a path to success in virtual screening. *Molecules* **2015**, *20*, 18732–18758.

(67) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(68) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general AMBER force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(69) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.

(70) Holmberg, N.; Ryde, U.; Bülow, L. Redesign of the coenzyme specificity in L-lactate dehydrogenase from Bacillus stearothermophilus using site-directed mutagenesis and media engineering. *Protein Eng.* **1999**, *12*, 851–856.

(71) Bryce, R. AMBER Parameter Database. `http://research.bmh.manchester.ac.uk/bryce/amber/`, Accessed: 2020-03-27.

(72) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132–146.

(73) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

(74) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*, 19–25.

(75) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; Van Gunsteren, W. F. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem. Phys. Lett.* **1994**, *222*, 529–539.

(76) Goga, N.; Rzepiela, A. J.; De Vries, A. H.; Marrink, S. J.; Berendsen, H. J. C. Efficient algorithms for Langevin and DPD dynamics. *J. Chem. Theory Comput.* **2012**, *8*, 3637–3649.

(77) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(78) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(79) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(80) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: an Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(81) Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976**, *22*, 245–268.

(82) Beauchamp, K. A.; Lim, N. M.; Mobley, D. L. Alchemical Analysis. `https://github.com/MobleyLab/alchemical-analysis`, 2015; Accessed: 2020-03-27.

(83) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129*, 124105.

(84) Mann, H. B.; Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **1947**, 50–60.

(85) Kruskal, W. H.; Wallis, W. A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621.

(86) Kendall, M. G. A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93.

(87) Kuhn, M.; Firth-Clark, S.; Tosco, P.; Mey, A. S.; Mackey, M. D.; Michel, J. Assessment of binding affinity via alchemical free energy calculations. *J. Chem. Inf. Model.* **2020**, *60*, 3120–3130.

(88) Bhakat, S.; Åberg, E.; Söderhjelm, P. Prediction of binding poses to FXR using multi-targeted docking combined with molecular dynamics and enhanced sampling. *J. Comput. Aided Mol. Des.* **2018**, *32*, 59–73.

(89) Wang, L.; Friesner, R. A.; Berne, B. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). *J. Phys. Chem. B.* **2011**, *115*, 9431–9438.

(90) Saha, A.; Shih, A. Y.; Mirzadegan, T.; Seierstad, M. Predicting the binding of fatty acid amide hydrolase inhibitors by free energy perturbation. *J. Chem. Theory Comput.* **2018**, *14*, 5815–5822.

(91) Schindler, C.; Rippmann, F.; Kuhn, D. Relative binding affinity prediction of farnesoid X receptor in the D3R Grand Challenge 2 using FEP+. *J. Comput. Aided Mol. Des.* **2018**, *32*, 265–272.