

# Accuracy and Precision of Alchemical Relative Free Energy Predictions with and without Replica-Exchange

Shunzhou Wan, Gary Tresadern, Laura Pérez-Benito, Herman van Vlijmen, and Peter V. Coveney\*

**A systematic and statistically robust protocol is applied for the evaluation of free energy calculations with and without replica-exchange. The protocol is based on ensemble averaging to generate accurate assessments of the uncertainties in the predictions. Comparison is made between FEP+ and TIES—free energy perturbation and thermodynamic integration with enhanced sampling—the latter with and without the so-called “enhanced sampling” based on replica-exchange protocols. Standard TIES performs best for a reference set of targets and compounds; no benefits accrue from replica-exchange methods. Evaluation of FEP+ and TIES with REST—replica-exchange with solute tempering—reveals a systematic and significant underestimation of free energy differences in FEP+, which becomes increasingly large for long duration simulations, is confirmed by extensive analysis of previous publications, and raises a number of questions pertaining to the accuracy of the predictions with the REST technique not hitherto discussed.**

## 1. Introduction

Notwithstanding numerous false dawns, we are now approaching an era when rapid, reliable, and reproducible free energy predictions for ligand–protein binding are becoming available.<sup>[1]</sup> Advances in free energy calculations have been fostered by the integration of improved force fields, enhanced sampling

Dr. S. Wan, Prof. P. V. Coveney  
Centre for Computational Science, Department of Chemistry  
University College London  
London WC1H 0AJ, UK  
E-mail: p.v.coveney@ucl.ac.uk

Prof. P. V. Coveney  
Computational Science Laboratory  
Institute for Informatics  
Faculty of Science  
University of Amsterdam  
Amsterdam 1098XH, The Netherlands

Dr. G. Tresadern, Dr. L. Pérez-Benito, Prof. H. van Vlijmen  
Computational Chemistry, Janssen Research & Development  
Janssen Pharmaceutica N. V.  
Turnhoutseweg 30, B-2340 Beerse, Belgium

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/adts.201900195>

© 2019 The Authors. Published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/adts.201900195

methods and increased computer power. The quality of predictions is understood to be determined by two principal sources: one is of a systematic nature owing to the force fields selected, the system setup, the algorithms, and their implementation, the sampling methodology, the extent of configurational space sampling, and so on;<sup>[2]</sup> the other is associated with the intrinsic stochasticity of the molecular dynamics simulation method used to compute binding affinities.<sup>[3]</sup> The accuracy is dominated by the former, while the precision is mainly determined by the latter.

Protein force fields are now generally quite reliable, although ligand parameterization for new molecules can still pose issues for modelers. The system setup needs to capture essential chemical details and

cannot be overlooked; the challenge to model ionization and tautomeric states accurately in a molecular mechanics formalism is a limitation (this is distinct from actual force field errors).<sup>[1a]</sup> The extent of conformational sampling also determines the quality of the calculations, as it is not possible to sample the entire conformational space for a complex molecular system, the most relevant subspace for a given thermodynamic property needs to be well sampled to obtain a converged ensemble average. It should be noted, however, that an apparently converged average, with no obvious change in time after a given period, does not necessarily mean that the relevant conformational space has been adequately sampled.<sup>[2b]</sup>

The stochastic uncertainty in predictions emanating from molecular dynamics simulations originates in the intrinsically chaotic behavior of the trajectories, as these display extreme sensitivity to initial conditions.<sup>[3]</sup> Historically, and still to an overwhelming degree today, an assumption is made that a single “long time” average provides the ensemble average from which statistical mechanics delivers macroscopic averages in the thermodynamic limit. However, these “long time” averages show no convincing tendency to converge to ensemble averages, which today can often be determined directly on large supercomputers.<sup>[3,4]</sup>

Such uncertainty quantification (UQ) furnishes a statistical estimate of the reproducibility of results between theory/simulation and experiment, and from two or more theoretical methods. Making comparisons can never be done from “one off” molecular dynamics results; both experiment and theory are riddled with errors and uncertainties and the issue is how to

reliably compare them. This requires not only computational results to be reported with errors, but likewise experiments.<sup>[5]</sup>

The ensemble approach is, in the more traditional jargon of molecular dynamics, a highly effective way of “sampling.” It is also a standard procedure adopted for UQ.<sup>[6]</sup> There have been many other efforts to accelerate sampling as compared to the traditional one of a single MD run. One such method is called replica-exchange with solute tempering (REST2)<sup>[7]</sup> in which interactions of a chosen subsystem with its environment are scaled to enhance overall conformational sampling for the regions of interest. REST2 is a modified version of the original “replica-exchange with solute tempering” (REST1)<sup>[8]</sup>: in REST1, the entire solute biomolecule is heated up while in REST2 the hot region is restricted to a local region of the solute. In Schrödinger’s proprietary FEP+ package, the REST2 method is implemented but called REST. For the sake of simplicity and consistency, hereafter we will call the approach REST. REST has been implemented in several molecular simulation packages including NAMD,<sup>[9]</sup> Amber,<sup>[10]</sup> GROMACS<sup>[11]</sup> as well as within the FEP+<sup>[12]</sup> package. As we shall see, these methods do not necessarily perform more accurately than ensemble averaged calculations of free energies without such “enhanced sampling.” There are claims that REST improves the convergence of free energy computations,<sup>[13]</sup> and hence reduces the variations in the prediction. Our previous study,<sup>[5b]</sup> however, has shown that the calculated free energy differences from five replicas vary by up to 1.2 kcal mol<sup>-1</sup> from TIES calculations with REST employed. Established “enhanced sampling” methods also require ensemble averaging as the stochastic variation is intrinsic to all MD-based methods.<sup>[3,5b]</sup>

Recently at Janssen, FEP+ simulations have been applied to different drug discovery projects,<sup>[14]</sup> and at UCL we have studied various series of ligands bound to a given protein target<sup>[5b,15]</sup> as well as the same ligands to wild-type and mutant proteins.<sup>[16]</sup> We have introduced variants of TIES, which incorporate the “enhanced sampling” REST techniques including the popular free energy estimator MBAR (multistate Bennett acceptance ratio).<sup>[5b]</sup> We have also extended the TIES methodology to study relative binding affinities caused by protein mutations when bound to a ligand, a variant which we call TIES-PM.<sup>[16]</sup> With REST, TIES-PM can capture large conformational changes; for example, it generates correct free energy differences caused by the gatekeeper mutation occurring inside the ligand binding pocket of the FGFR-1 kinase.<sup>[16]</sup> Simulations using standard TIES, without REST, cannot overcome significant energy barriers between conformations and hence the results are highly sensitive to the initial structures. Nonetheless, we have also observed cases where the application of REST degrades the quality of free energy predictions.<sup>[16]</sup> These particular cases demand extensive evaluation with large datasets to take them out of the class of the anecdotal and into the domain of the scientific.

The purpose of the present paper is to assess the performance of “sampling” strategies in two approaches to relative free energy prediction, TIES with and without replica-exchange along with FEP+. As noted above, a recent paper reported a direct comparison between TIES computed using NAMD and pmemdGTI, where the same force field was used but with different protocols.<sup>[5b]</sup> Here we use the reference data set from previous publications<sup>[12,15a,c]</sup> to look at the predictions from the TIES and

**Table 1.** Molecular systems and simulation protocols.

	TIES	TIES- $\lambda$ -REST (-MBAR)	FEP+
Molecular systems	BRD4 (12 ligand pairs); MCL1 (16 ligand pairs); TYK2 (11 ligand pairs)		
Force field	Amber ff99SBildn		OPLS3e
MD engine	NAMD 2.9	NAMD 2.11 <sup>a)</sup>	Desmond v3.8.5.19
Topology	Dual topology		Single topology
Box buffer [Å]	14		10
Timestep [fs]	2		4/8
Cut-off [Å]	12		9
$\lambda$ windows	13		12
Protocol	5 replicas, 4 ns production run for each $\lambda$ window		
Extension <sup>b)</sup>	–	4 ns, 20 replicas 40 ns, 10 replicas	4 ns, 30 replicas 40 ns, 20 replicas
Hours/ns <sup>c)</sup>	1.41 <sup>d)</sup>	2.36 <sup>e)</sup>	0.05 <sup>f)</sup>

<sup>a)</sup> Customized version of the NAMD 2.11 package with a patch to implement the REST algorithm for alchemical simulations; <sup>b)</sup> 5 ligand pairs are chosen for each protein system; <sup>c)</sup> use TYK2 as an example; <sup>d)</sup> 128 cores on SuperMUC for one  $\lambda$  window; <sup>e)</sup> 124 cores on BlueWaters for one  $\lambda$  window; <sup>f)</sup> 4x Nvidia Tesla K80.

FEP+ protocols. The paper is structured as follows: in the next section, we lay out the methods used; in the following one, we present the results and a discussion. The paper ends with our conclusions from the study.

## 2. Experimental Section

The molecular systems and the simulation protocols are summarized in **Table 1**. Three molecular systems were used: BRD4 (bromodomain containing 4),<sup>[15c]</sup> MCL1 (myeloid cell leukemia 1), and TYK2 (Tyrosine kinase 2).<sup>[12,15a]</sup> The same molecular systems as used before with the TIES approach were studied,<sup>[15a,c]</sup> allowing comparison of the accuracy of normal TIES with FEP+ and TIES with REST (TIES- $\lambda$ -REST). The same initial structures and same ligand pairs as in previous studies<sup>[15a,c]</sup> were used for both FEP+ and TIES- $\lambda$ -REST simulations. The simulation-ready molecular systems, in Amber format for TIES and maegz format for FEP+, can be found in Supporting Information.

### 2.1. TIES- $\lambda$ -REST Simulations

For the purpose of comparison, the same Amber ff99SBildn<sup>[17]</sup> force field was used in TIES- $\lambda$ -REST as in TIES calculations.<sup>[15a,c]</sup> The same procedures to set up the protein–ligand systems as recently reported and validated were used.<sup>[5b]</sup> A customized version of the NAMD 2.11 package,<sup>[18]</sup> with implementation of REST for alchemical simulations,<sup>[9]</sup> was used for all the TIES- $\lambda$ -REST simulations. A dual topology scheme was employed to describe a hybrid ligand which consisted of both a disappearing and an appearing group. The two groups define all the alchemically mutating atoms, exclusively belonging to the two ligands which are transferred from one to another during the alchemical process. Thermodynamic integration was used to calculate the free energy changes  $\Delta G_{\text{alch}}$  for the ligand pairs in protein and in solvent.

The binding free energy difference  $\Delta\Delta G_{\text{cal}}$  was then calculated as the differences of the  $\Delta G_{\text{alch}}$  values from the two simulations. The REST region for unbound ligand calculations was defined as the set of alchemically mutating atoms. For bound ligand calculations, the REST region comprised all alchemically mutating atoms and all protein residues within 3 Å distance of the former. All TIES( $\lambda$ -REST) simulations use 13  $\lambda$  windows per perturbation. A soft core potential was applied for the van der Waals interactions of all atoms in the alchemical space. No soft-core potential was used for the electrostatic interactions. For the disappearing atoms, the electrostatic interactions were linearly decoupled from the simulations between  $\lambda$  values of 0 and 0.55 and completely turned off beyond that; for the appearing atoms, they were linearly coupled to the simulations from  $\lambda$  value 0.45–1, and fully extinguished otherwise. The approach of decoupling/coupling at different rates ensured that the partial charges were removed on perturbed atoms before they were fully annihilated, while the charges on the growing atoms were introduced after they appeared.

Each REST simulation involves running a predefined number of parallel REST replicas, 13 in the TIES- $\lambda$ -REST simulations, varying in both their effective temperatures and the alchemical parameter  $\lambda$ .<sup>[16]</sup> Regular exchange of configurations was attempted between neighboring REST replicas. All TIES- $\lambda$ -REST simulations were run on the BlueWaters supercomputer at the National Center for Supercomputing Applications of the University of Illinois at Urbana–Champaign and Titan at Oak Ridge National Laboratory. The previous TIES calculations were run on the SuperMUC Phase 1 and 2 computers at the Leibniz Supercomputing Centre (Table 1). The benchmark simulations showed that TIES- $\lambda$ -REST consumed about 10–20% more node hours than the standard TIES approach.

## 2.2. FEP+ Approach

FEP+ calculations were performed using Maestro v2018.2, Desmond multisim version 3.8.5.19 and mmshare version 4.2, along with the first version of the proprietary OPLS3e force field. A recent OPLS3e paper<sup>[19]</sup> showed that some changes had been made since OPLS3, one being replacement of a quantum chemical MP2 calculation with a density functional method for the torsions, another being an extended chemical environment for torsion definition. A REST enhanced sampling technique was used in the Desmond MD engine, with the same effect as REST in TIES- $\lambda$ -REST simulations. The default FEP+ protocol was used to define the REST region in which only perturbed ligand atoms were included for the simulations in water and in protein complexes (note that a different REST region was used in TIES- $\lambda$ -REST for complex simulations). Missing force field parameters were added by additional QM calculations<sup>[19]</sup> and fitted using the fbuilder module. A single topology scheme was used, in which corresponding atoms were mapped between the two ligands via a maximum common substructure search. No cycle-closure averaging and error estimation were used as the free energy changes were calculated individually for each ligand pair. FEP+ uses 12  $\lambda$  windows per perturbation in both solvent and complex, and makes use of a mixed coupling/decoupling schedule. Bonded interactions were scaled linearly across all 12  $\lambda$  windows as they

were either removed or introduced. Desmond used soft core potentials to overcome possible van der Waals end point instabilities at the limits of the  $\lambda$  coordinate. For atom deletion, charges were decoupled first in a linear manner during the first five  $\lambda$  windows with the remaining seven  $\lambda$  windows used to turn-off the van der Waals terms. The reverse schedule was used for introducing atoms. There are other settings in FEP+, which differ from those in TIES simulations; these included a smaller box size, a smaller cut-off distance for non-bonded interactions, and larger time steps (Table 1). All FEP+ Desmond runs were performed on an in-house GPU (Nvidia Tesla K80) cluster at Janssen Research & Development, Beerse, Belgium.

## 2.3. Simulation Protocol

For both TIES- $\lambda$ -REST and FEP+ simulations, the protocol established in the previous publications was used, in which an ensemble of five replicas had been employed.<sup>[5b,a]</sup> It should be noted that such ensemble-based simulation is not part of the “standard” FEP+ protocol. All replicas have identical initial coordinates but different velocities drawn randomly from a Maxwell–Boltzmann distribution. 4 ns production runs were performed for each replica.<sup>[15a]</sup> The protocol of five replicas and 4 ns production times had been shown to produce accurate, precise, and reliable relative free energies in the previous TIES studies on various molecular systems,<sup>[5b,15,16]</sup> of which a subset was invoked to make a direct comparison between approaches with the replica-exchange method, as well as between its use and non-use. The MBAR approach was automatically applied in FEP+ simulations to generate free energy estimates. For TIES- $\lambda$ -REST, the free energy differences both with and without MBAR were reported. The protocol of five replicas and 4 ns production runs were established using TIES, with or without REST.<sup>[5b,15a]</sup> To test this, in the unlikely case that a different combination might favor FEP+, the number of replicas and the duration of the production MD runs were also varied. Thus, some simulations were extended up to 30 replicas and 40 ns (Table 1). The accuracy of free energy approaches was therefore evaluated by comparing the calculations with the experimental data using mean signed errors (MSEs), mean unsigned errors (MUEs), root-mean-square errors (RMSEs), and linear regression. For the experimental data without uncertainties reported, an approximate error of 0.3 log units (0.41 kcal mol<sup>-1</sup>) was used.<sup>[20]</sup> The way TIES computes precisions was from the TI integral correctly interpreted as a stochastic integral.<sup>[3,15a]</sup> For FEP+, the error was provided for each individual replica according to standard MBAR theory. The MBAR errors, however, largely underestimate the variances of the free energy results from replica calculations (see Section 3). To avoid such issues and make the performance of the TIES-based approaches and FEP+ comparable, the standard deviations were used here to assess the precision for all of the approaches.

## 3. Results and Discussion

In this section, we compare the accuracy and precision of free energy estimates from different approaches: TIES, TIES- $\lambda$ -REST

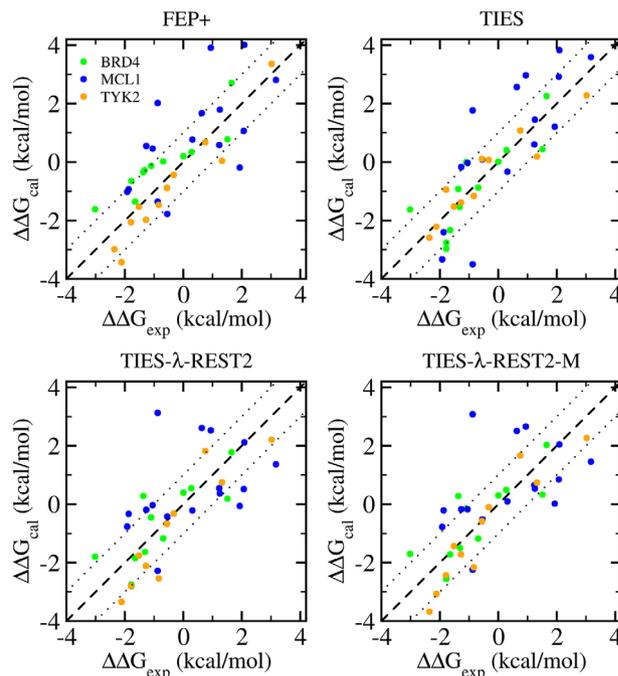
**Table 2.** Free energy predictions from FEP+, original TIES, TIES- $\lambda$ -REST, and TIES- $\lambda$ -REST-MBAR. All the results are from ensemble simulations consisting of five replicas and 4 ns production runs each. Standard deviations in parentheses.

Protein	Property	FEP+	TIES	TIES- $\lambda$ -REST	TIES- $\lambda$ -REST-M
BRD4	No. of pert.			12	
	MUE	0.81 (0.14)	0.67 (0.15)	0.72 (0.16)	0.69 (0.15)
	MSE	-0.59 (0.22)	0.00 (0.25)	-0.09 (0.27)	-0.16 (0.25)
	RMSE	0.91 (0.14)	0.81 (0.16)	0.87 (0.17)	0.84 (0.16)
	Pearson r	0.90	0.84	0.80	0.81
	Slope	0.72	0.92	0.82	0.81
	intercept	0.47	-0.11	-0.13	-0.03
MCL1	No. of pert.			16	
	MUE	1.30 (0.24)	1.20 (0.23)	1.34 (0.28)	1.26 (0.25)
	MSE	-0.22 (0.42)	0.27 (0.39)	-0.71 (0.43)	-0.64 (0.39)
	RMSE	1.53 (0.27)	1.41 (0.26)	1.61 (0.33)	1.56 (0.31)
	Pearson r	0.61	0.80	0.41	0.44
	Slope	0.68	1.18	0.37	0.40
	intercept	0.68	0.32	0.44	0.48
TYK2	No. of pert.			11	
	MUE	0.51 (0.16)	0.44 (0.15)	0.85 (0.21)	0.66 (0.17)
	MSE	0.27 (0.26)	-0.25 (0.23)	0.60 (0.32)	0.36 (0.28)
	RMSE	0.67 (0.19)	0.56 (0.17)	1.02 (0.23)	0.79 (0.19)
	Pearson r	0.97	0.94	0.93	0.94
	Slope	1.11	0.83	1.20	1.11
	intercept	-0.39	-0.12	-0.55	-0.38
All	No. of pert.			39	
	MUE	0.93 (0.13)	0.82 (0.12)	1.01 (0.15)	0.91 (0.13)
	MSE	-0.20 (0.20)	0.04 (0.19)	-0.15 (0.23)	-0.21 (0.21)
	RMSE	1.16 (0.16)	1.05 (0.16)	1.26 (0.19)	1.18 (0.18)
	Pearson r	0.78	0.84	0.72	0.74
	Slope	0.87	1.04	0.81	0.79
	intercept	0.29	0.14	-0.13	-0.02

Mean unsigned error  $MUE = \sum_{i=1}^n \frac{|\Delta\Delta G_{\text{exp}} - \Delta\Delta G_{\text{cal}}|}{n}$ ; mean signed error  $MSE = \sum_{i=1}^n \frac{\Delta\Delta G_{\text{cal}} - \Delta\Delta G_{\text{exp}}}{n}$  (rearrange each ligand pair so that  $\Delta\Delta G_{\text{exp}} \geq 0$ ); root mean squared error  $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\Delta\Delta G_{\text{exp}} - \Delta\Delta G_{\text{cal}})^2}$ . Slope ( $m$ ) and intercept ( $b$ ) are defined as in  $\Delta\Delta G_{\text{cal}} = \alpha * \Delta\Delta G_{\text{exp}} + \beta$ .

(-MBAR), and FEP+. Between TIES- $\lambda$ -REST and TIES- $\lambda$ -REST-MBAR, we consider the former for reasons of simplification, as the two approaches generate very similar results in terms of MSEs, MUEs, RMSEs, and correlation coefficients (Table 2). Our previous studies have shown that MBAR does little to enhance such free energy predictions.<sup>[5b,16]</sup> All approaches generate good predictions when compared with the experimental data (Figure 1 and Table 2). Taking together all of the quantifications in Table 2, it can be seen that TIES yields the best results, with the smallest MSEs, MUEs, and RMSEs and the best correlations in most cases. It should be said that these statistical differences are only marginal. The predictions from FEP+ are comparable to the results from TIES- $\lambda$ -REST.

While ensemble approaches diminish random errors from simulations, they cannot remove systematic bias, a deviation of



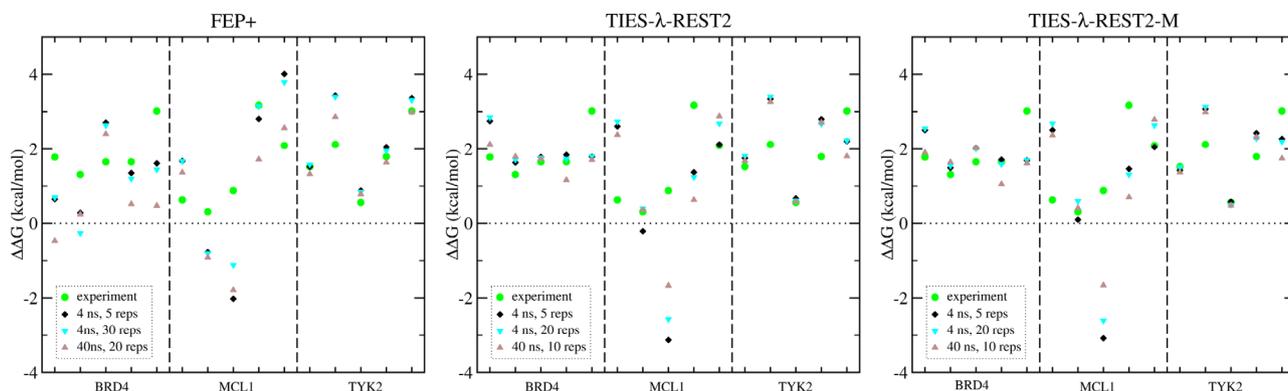
**Figure 1.** Comparison of the predicted binding free energy differences with the experimental data from the four approaches. See Table 2 for the errors and correlations.

a measurement or prediction from the true value. Indeed, recent work shows that even ensemble averages are likely to contain systematic errors, caused by a newly discovered pathology of floating point numbers.<sup>[21]</sup> The possible bias for each of the above approach is indicated by the derivation of the slope from one in the linear regression  $\Delta\Delta G_{\text{cal}} = \alpha * \Delta\Delta G_{\text{exp}} + \beta$ , and can be quantified by the MSE values (Table 2). An overestimation is defined by a positive MSE,  $(\Delta\Delta G_{\text{cal}} - \Delta\Delta G_{\text{exp}}) > 0$  and a slope greater than one when ligand pairs are rearranged so that  $\Delta\Delta G_{\text{exp}} > 0$ ; conversely, an underestimation arises when  $(\Delta\Delta G_{\text{cal}} - \Delta\Delta G_{\text{exp}}) < 0$  and the slope  $< 1$ . TIES results exhibit negligible bias with small MSEs for each protein system and for the entire data set. Indeed, TIES generates an average overestimation of  $0.27 \text{ kcal mol}^{-1}$  for MCL1, an underestimation of  $0.25 \text{ kcal mol}^{-1}$  for TYK2, no bias ( $0 \text{ kcal mol}^{-1}$ ) for BRD4, and no bias ( $0.04 \text{ kcal mol}^{-1}$ ) for the entire dataset; the slope for the fitted line is also close to one for the entire dataset. Conversely, FEP+ manifests a bias in the case of BRD4 with  $0.59 \text{ kcal mol}^{-1}$  underestimation, while generating equal but opposite MSEs for the other two systems. For the entire data set, FEP+ underestimated the relative free energy changes by  $0.20 \text{ kcal mol}^{-1}$ . TIES- $\lambda$ -REST shows similar bias for each perturbation as FEP+ but with different magnitudes, and a similar underestimation for the entire dataset. It should be noted, however, that the dataset used here is still relatively small; a large and broad dataset with multi-target and multi chemotype will be required to have more statistically significant conclusions.

Using more replicas does not confer significant benefit on the predictions of binding free energy differences in either FEP+ or TIES- $\lambda$ -REST (-MBAR); MUEs, MSEs, RMSEs, and correlation coefficients are all comparable between the predictions from simulations with five replicas and 20–30 replicas (Table 3). Likewise,

**Table 3.** Results of FEP+ and TIES- $\lambda$ -REST(-MBAR) with different numbers of replicas and simulation lengths. Standard deviations in parentheses.

Protein	Property	FEP+			TIES- $\lambda$ -REST			TIES- $\lambda$ -REST-MBAR		
		4 ns 5 reps	4 ns 30 reps	40 ns 20 reps	4 ns 5 reps	4 ns 20 reps	40 ns 10 reps	4 ns 5 reps	4 ns 20 reps	40 ns 10 reps
BRD4	No. of pert.	15								
MCL1	MUE	0.96 (0.21)	0.91 (0.18)	1.04 (0.22)	0.97 (0.28)	0.96 (0.27)	0.92 (0.23)	0.86 (0.28)	0.89 (0.26)	0.89 (0.23)
TYK2	MSE	-0.13 (0.35)	-0.13 (0.32)	-0.65 (0.33)	-0.15 (0.40)	-0.03 (0.40)	-0.16 (0.36)	-0.22 (0.38)	-0.12 (0.38)	-0.25 (0.24)
	RMSE	1.21 (0.26)	1.10 (0.21)	1.33 (0.25)	1.40 (0.39)	1.32 (0.35)	1.23 (0.27)	1.33 (0.39)	1.28 (0.36)	1.19 (0.27)
	Correlation	0.87	0.89	0.78	0.77	0.79	0.79	0.78	0.80	0.80
	Slope	1.01	1.00	0.69	0.87	0.91	0.81	0.85	0.87	0.77
	intercept	0.52	0.52	-0.48	-0.02	0.03	-0.03	0.12	0.17	0.08



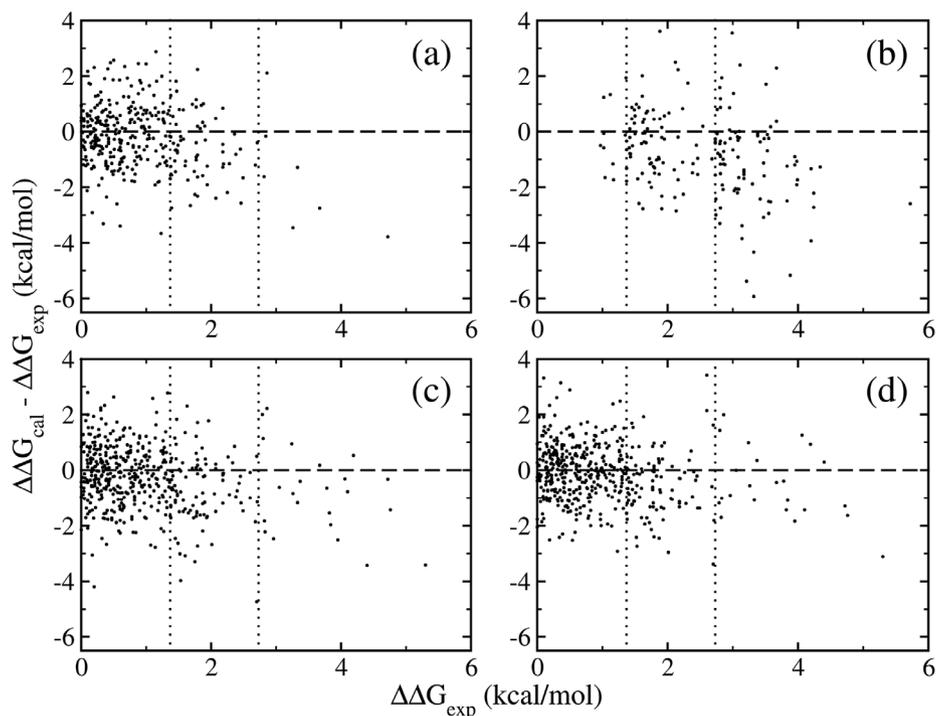
**Figure 2.** Changes of predicted binding free energies when the number of replicas is increased and/or the length of simulations is extended.

for predictions using standard TIES<sup>[5b]</sup> and the TIES- $\lambda$ -REST approach, longer simulation time does not make a significant impact on the predictions, based on the values of MUEs, MSEs, RMSEs, and correlation coefficients (Table 3 and Figure 2, confirming our earlier findings). However, and by contrast, in the FEP+ approach longer simulation time degrades performance, which is discussed further below. It renders the underestimation even more severe, as indicated by the MSE values that increase from 0.13 to 0.65 kcal mol<sup>-1</sup> for the subset of ligand pairs for which the simulation duration is increased by an order of magnitude (Table 3).

The underestimation of FEP+ relative free energy calculations has been recently noted.<sup>[20]</sup> To further quantify the underestimation in FEP+ calculations, we revisit three sets of predictions from two Schrödinger publications<sup>[12,19]</sup> and one set from the recent study by Pérez-Benito et al.<sup>[20]</sup> (see Figure 3 and Table 4). These studies all looked at relatively large datasets and thus produce reliable statistics. The original  $\Delta\Delta G_{\text{FEP+}}$  values were obtained from Supporting Information<sup>[12,19]</sup> and from the authors.<sup>[20]</sup> In the original FEP+ paper with OPLS2.1 force field,<sup>[12]</sup> a 0.18 kcal mol<sup>-1</sup> underestimation is manifest for the entire dataset consisting of 330 alchemical mutations. Using the latest versions of the OPLS force field, similar underestimations have been observed.<sup>[19]</sup> These underestimations are comparable with the MSE value (0.20 kcal mol<sup>-1</sup>, Table 2) in the current study. The underestimation in FEP+ is much more obvious and severe when the binding free energy differences are large. When the  $\Delta\Delta G_{\text{exp}}$  values are in the range of 1.37–2.73 kcal mol<sup>-1</sup>, 1–2 log

units in activity, the underestimation is  $\approx 0.50$  kcal mol<sup>-1</sup>; when the difference is more than 2 log units (2.73 kcal mol<sup>-1</sup>), the underestimation can be as much as 1.56 kcal mol<sup>-1</sup> (Table 4). The comparison of the results from 1 and 5 ns simulations<sup>[20]</sup> shows that long simulations degrade the quality of FEP+ predictions; longer simulations make them even worse, as observed in the current study (Table 3).

REST has been implemented in FEP+ and TIES- $\lambda$ -REST approaches. A larger database will be required to confirm the apparent underestimations of the free energy differences in TIES- $\lambda$ -REST-MBAR. Although the quality of these simulation results may be affected by many factors,<sup>[2b]</sup> we suspect that the REST protocol is the dominant reason for the underestimations exhibited in these simulations. TIES and TIES- $\lambda$ -REST(-MBAR) simulations share the same protocol including the force field and the initial structures. The only difference is the use of REST in TIES- $\lambda$ -REST(-MBAR), which is likely to be the reason that TIES calculations outperform TIES- $\lambda$ -REST(-MBAR) in its relative free energy predictions. The REST approach enhances conformational sampling, but by the same token is able to reach conformations which are less relevant to stable binding and can produce diminished differences in binding affinities for pairs of congeneric ligands, as recently reported.<sup>[16]</sup> The lack of the correct weighting of these less relevant states in REST-implemented calculations causes an artefactual reduction in the difference of the binding free energies.<sup>[16]</sup> Longer REST simulations increase the occurrence of such conformations and hence make the predictions increasingly unreliable. For the FEP+



**Figure 3.** Underestimation of the relative free energy differences. Data are taken from previous FEP+ calculations from a) Wang et al.,<sup>[12]</sup> b) Pérez-Benito et al.,<sup>[20]</sup> c) Roos et al.<sup>[19]</sup> using OPLS3, and d) Roos et al.<sup>[19]</sup> using OPLS3e force field. See Table 4 for quantitative assessments.

**Table 4.** Revisit binding free energy differences ( $\text{kcal mol}^{-1}$ ) of FEP+ calculations in literature.

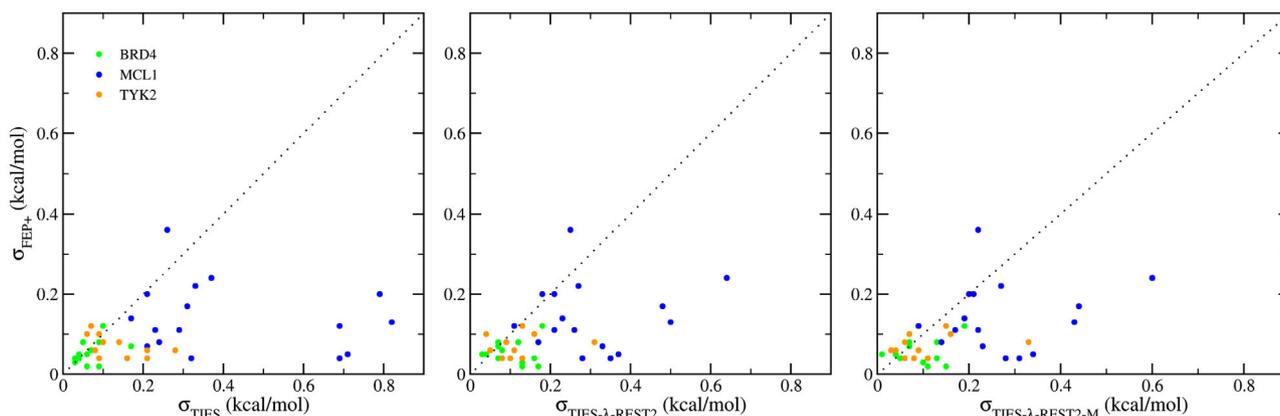
	$ \Delta\Delta G_{\text{exp}}  < 1.37$	$1.37 \leq  \Delta\Delta G_{\text{exp}}  < 2.73$	$ \Delta\Delta G_{\text{exp}}  \geq 2.73$	Total
Wang et al., <sup>[12]</sup> OPLS2.1, Figure 3a				
No. of transformations	257	66	7	330
Underestimations [%]	51	71	86	55
$(\Delta\Delta G_{\text{cal}} - \Delta\Delta G_{\text{exp}})/N$	-0.04	-0.59	-1.56	-0.18
Roos et al., <sup>[19]</sup> OPLS3, Figure 3c				
No. of transformations	348	98	23	469
Underestimations [%]	55	71	74	59
$(\Delta\Delta G_{\text{cal}} - \Delta\Delta G_{\text{exp}})/N$	-0.12	-0.71	-0.82	-0.28
Roos et al., <sup>[19]</sup> OPLS3e, Figure 3d				
No. of transformations	348	98	23	469
Underestimations [%]	52	71	65	57
$(\Delta\Delta G_{\text{cal}} - \Delta\Delta G_{\text{exp}})/N$	-0.03	-0.53	-0.51	-0.16
Pérez-Benito et al., <sup>[20]</sup> all ligand pairs with 5 ns simulations from LO datasets, Figure 3b				
No. of transformations	10	79	92	181
Underestimations [%]	70	65	83	74
$(\Delta\Delta G_{\text{cal}} - \Delta\Delta G_{\text{exp}})/N$	-0.13	-0.43	-1.14	-0.78

"Underestimations" is the percentage of pairs of which  $\Delta\Delta G_{\text{cal}} < \Delta\Delta G_{\text{exp}}$  when the ligand pairs are rearranged so that  $\Delta\Delta G_{\text{exp}} \geq 0 \text{ kcal mol}^{-1}$ . The average underestimation  $(\Delta\Delta G_{\text{cal}} - \Delta\Delta G_{\text{exp}})/N$  is also calculated after making the  $\Delta\Delta G_{\text{exp}} \geq 0$  rearrangement.

calculations, the force field may also contribute to the bias, as the latest OPLS3e force field reduces the underestimations in the free energy predictions as compared with OPLS3 force field (Table 2 and ref. [19]).

Although FEP+ manifests smaller variations between different replicas than TIES or TIES- $\lambda$ -REST(MBAR) (Figure 4), it

produces much larger variations from 30 replicas than the MBAR errors reported for each FEP+ calculation (Table 5). The results from single FEP+ calculations can vary by up to  $3.94 \text{ kcal mol}^{-1}$  for one ligand pair to MCL1. Although the implementation of an accelerated sampling protocol such as REST may reduce the variations from independent runs, the results here show again that



**Figure 4.** Comparison of the bootstrapping errors of FEP+ with TIES, TIES- $\lambda$ -REST, and TIES- $\lambda$ -REST-MBAR.

**Table 5.** Variations and ranges of binding free energy differences (kcal mol<sup>-1</sup>), compared with the averaged MBAR errors, from 30 replicas of 4-ns FEP+ calculations. Standard deviations in parentheses.

Protein	Ligand pair	Average	Range from 30 replicas	$\sigma_{\text{MBAR}}$
BRD4	02–03	−2.64 (0.15)	−3.02 to −2.38 (0.64)	0.08
	03–06	−1.20 (0.27)	−1.80 to −0.69 (1.11)	0.12
	08–09	1.44 (0.16)	0.98 to 1.71 (0.73)	0.09
	13–14	−0.70 (0.30)	−1.35 to −0.12 (1.23)	0.11
	14–15	0.26 (0.10)	0.14 to 0.47 (0.33)	0.04
MCL1	02–32	−1.11 (0.95)	−3.44 to 0.50 (3.94)	0.31
	08–18	−3.79 (0.23)	−4.21 to −3.09 (1.12)	0.07
	16–34	0.80 (0.11)	0.53 to 1.06 (0.53)	0.05
	32–38	−3.15 (0.54)	−4.30 to −2.09 (2.21)	0.19
	35–12	1.67 (0.34)	0.95 to 2.27 (1.32)	0.07
TYK2	01–03	0.85 (0.28)	0.36 to 1.58 (1.22)	0.12
	01–08	1.96 (0.17)	1.49 to 2.22 (0.73)	0.10
	01–10	3.40 (0.23)	2.90 to 3.89 (0.99)	0.11
	06–11	−3.31 (0.17)	−3.79 to −2.90 (0.89)	0.08
	06–15	−1.58 (0.17)	−1.94 to −1.21 (0.73)	0.08

one-off runs are not reliable. Statistical properties derived from ensembles are much more robust.

It is likely that the number of atoms in the alchemical region plays an important role in accounting for the differences of errors from TIES and FEP+ (Figure 4). FEP+ implements a single topology approach which morphs as many atoms as possible that are different between the two end states. TIES and its variants use a dual topology approach in which the appearing and disappearing groups are not morphed and move independently. If we define dummy atoms as those having no interactions with the environment at one or the other end state, the number of dummy atoms is usually significantly larger in a dual topology than in a single topology scheme. These dummy atoms can sample different conformational spaces and may be very flexible, especially when their interactions with the environment are scaled down. This introduces larger variations in the energies and energy derivatives in TIES and TIES- $\lambda$ -REST(-MBAR) than FEP+.

## 4. Conclusion

In this study, we compare the accuracy and precision of relative free energies calculated from standard TIES<sup>[15a,c]</sup> and two REST-implemented approaches: TIES- $\lambda$ -REST (with or without MBAR)<sup>[5b]</sup> and FEP+.<sup>[12]</sup> The performance of standard TIES has been compared with pmemdGTI in our previous study.<sup>[5b]</sup> For the reference data set investigated, standard TIES performs best. The protocol of a 4 ns production run and five replicas, as established in our previous TIES studies,<sup>[5b,15,16]</sup> is reasonable for TIES and its variants, since neither an increase in the number of replicas nor the duration of simulations have a large impact on the predictions, as adjudged by the quantities MUEs, MSEs, RMSEs, and correlation coefficients (Figure 2 and Table 3). Indeed, an interesting conclusion from this study is that more replicas does not confer significant benefit on the predictions of binding free energy differences. However, FEP+ accuracy deteriorates as the simulation duration is extended.

The REST-implemented calculations all show an underestimation of computed relative free energies, which are especially noticeable when the experimental binding free energy difference is large. Longer simulations degrade predictions in FEP+ when conformations are sampled which are less relevant to stable ligand binding. Proper weighting is required for the entire conformational space sampled to produce reliable free energy predictions. However, it is difficult to evaluate the likelihood of a conformation and hence its contribution to the predictions.<sup>[16]</sup> Force fields can also contribute to the observed bias in relative free energy predictions; the latest OPLS3e force field improves the FEP+ results but does not remove its systematic underestimations.

## Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

## Acknowledgements

The authors thank Dr. Agastya P. Bhati for helping in selecting the ligand pairs used in this study, and Dr. David W. Wright for valuable discussions. The authors would like to acknowledge i) the UK Medical Research Council for funding the Medical Bioinformatics project (MR/L016311/1), ii)

Qatar National Research Fund (Grant No. 7-1083-1-191), iii) EU H2020 projects ComPat (<http://www.compat-project.eu/>, Grant No. 671564), CompBioMed and CompBioMed2 (<http://www.compbioed.eu>, Grant Nos 675451 and 823712), iv) NSF Award (<https://www.nsf.gov/pubs/2017/nsf17542/nsf17542.htm>, Award No. NSF 1713749) and v) special funding to PVC from the UCL Provost. The authors made use of the BlueWaters supercomputer at the National Center for Supercomputing Applications of the University of Illinois at Urbana-Champaign (<https://bluewaters.ncsa.illinois.edu>), access to which was made available through the aforementioned NSF award, and the Titan supercomputer at the Oak Ridge National Laboratory, supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

## Conflict of Interest

The authors declare no conflict of interest.

## Keywords

binding free energies, FEP+, replica-exchange, thermodynamic integration with enhanced sampling

Received: October 2, 2019

Revised: October 30, 2019

Published online:

- [1] a) B. Sherborne, V. Shanmugasundaram, A. C. Cheng, C. D. Christ, R. L. Desjarlais, J. S. Duca, R. A. Lewis, D. A. Loughney, E. S. Manas, G. B. McGaughey, C. E. Peishoff, H. van Vlijmen, *J. Comput.-Aided Mol. Des.* **2016**, *30*, 1139; b) B. J. Williams-Noonan, E. Yuriev, D. K. Chalmers, *J. Med. Chem.* **2018**, *61*, 638.
- [2] a) J. D. Chodera, D. L. Mobley, M. R. Shirts, R. W. Dixon, K. Branson, V. S. Pande, *Curr. Opin. Struct. Biol.* **2011**, *21*, 150; b) W. F. van Gunsteren, X. Daura, N. Hansen, A. E. Mark, C. Oostenbrink, S. Riniker, L. J. Smith, *Angew. Chem., Int. Ed.* **2018**, *57*, 884.
- [3] P. V. Coveney, S. Wan, *Phys. Chem. Chem. Phys.* **2016**, *18*, 30236.
- [4] S. Wan, B. Knapp, D. W. Wright, C. M. Deane, P. V. Coveney, *J. Chem. Theory Comput.* **2015**, *11*, 3346.
- [5] a) T. D. Bunney, S. Wan, N. Thiyagarajan, L. Sutto, S. V. Williams, P. Ashford, H. Koss, M. A. Knowles, F. L. Gervasio, P. V. Coveney, M. Katan, *EBioMedicine* **2015**, *2*, 194; b) A. P. Bhati, S. Wan, Y. Hu, B. Sherborne, P. V. Coveney, *J. Chem. Theory Comput.* **2018**, *14*, 2867.
- [6] a) VECMA, <https://www.vecma.eu/> (accessed: October 2019); b) VECMAtk, <https://www.vecma-toolkit.eu/> (accessed: October 2019).
- [7] L. Wang, R. A. Friesner, B. J. Berne, *J. Phys. Chem. B* **2011**, *115*, 9431.
- [8] P. Liu, B. Kim, R. A. Friesner, B. J. Berne, *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 13749.
- [9] S. Jo, W. Jiang, *Comput. Phys. Commun.* **2015**, *197*, 304.
- [10] D. J. Mermelstein, C. Lin, G. Nelson, R. Kretsch, J. A. McCammon, R. C. Walker, *J. Comput. Chem.* **2018**, *39*, 1354.
- [11] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, E. Lindahl, *SoftwareX* **2015**, *1–2*, 19.
- [12] L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, D. L. Romero, C. Masse, J. L. Knight, T. Steinbrecher, T. Beuming, W. Damm, E. Harder, W. Sherman, M. Brewer, R. Wester, M. Murcko, L. Frye, R. Farid, T. Lin, D. L. Mobley, W. L. Jorgensen, B. J. Berne, R. A. Friesner, R. Abel, *J. Am. Chem. Soc.* **2015**, *137*, 2695.
- [13] W. Jiang, B. Roux, *J. Chem. Theory Comput.* **2010**, *6*, 2559.
- [14] a) L. Pérez-Benito, H. Keränen, H. van Vlijmen, G. Tresadern, *Sci. Rep.* **2018**, *8*, 4883; b) H. Keränen, L. Pérez-Benito, M. Ciordia, F. Delgado, T. B. Steinbrecher, D. Oehlrich, H. W. T. van Vlijmen, A. A. Trabanco, G. Tresadern, *J. Chem. Theory Comput.* **2017**, *13*, 1439; c) M. Ciordia, L. Pérez-Benito, F. Delgado, A. A. Trabanco, G. Tresadern, *J. Chem. Inf. Model.* **2016**, *56*, 1856; d) F. J. R. Rombouts, G. Tresadern, P. Buijnsters, X. Langlois, F. Tovar, T. B. Steinbrecher, G. Vanhoof, M. Somers, J.-I. Andrés, A. A. Trabanco, *ACS Med. Chem. Lett.* **2015**, *6*, 282.
- [15] a) A. P. Bhati, S. Wan, D. W. Wright, P. V. Coveney, *J. Chem. Theory Comput.* **2017**, *13*, 210; b) S. Wan, A. P. Bhati, S. Skerratt, K. Omoto, V. Shanmugasundaram, S. K. Bagal, P. V. Coveney, *J. Chem. Inf. Model.* **2017**, *57*, 897; c) S. Wan, A. P. Bhati, S. J. Zasada, I. Wall, D. Green, P. Bamborough, P. V. Coveney, *J. Chem. Theory Comput.* **2017**, *13*, 784.
- [16] A. P. Bhati, S. Wan, P. V. Coveney, *J. Chem. Theory Comput.* **2019**, *15*, 1265.
- [17] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, D. E. Shaw, *Proteins* **2010**, *78*, 1950.
- [18] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, K. Schulten, *J. Comput. Chem.* **2005**, *26*, 1781.
- [19] K. Roos, C. Wu, W. Damm, M. Reboul, J. M. Stevenson, C. Lu, M. K. Dahlgren, S. Mondal, W. Chen, L. Wang, R. Abel, R. A. Friesner, E. D. Harder, *J. Chem. Theory Comput.* **2019**, *15*, 1863.
- [20] L. Pérez-Benito, N. Casajuana-Martin, M. Jimenez-Roses, H. van Vlijmen, G. Tresadern, *J. Chem. Theory Comput.* **2019**, *15*, 1884.
- [21] B. M. Boghosian, P. V. Coveney, H. Wang, *Adv. Theory Simul.* **2019**, <https://doi.org/10.1002/adts.201900125>.