Evaluation and Characterization of Isoxazole Amides as SMYD3 Inhibitors

Shunzhou Wan[1], Agastya Bhati[1], David Wright[1],

Ian Wall[2], Alan Graves[2], Darren Green[2], Peter V. Coveney[1,3,4*]

[1]Centre for Computational Science, Department of Chemistry, University College London, London WC1H 0AJ, United Kingdom
[2]GlaxoSmithKline, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, United Kingdom
[3]Advanced Research Computing Centre, University College London, London WC1H 0AJ, United Kingdom
[4]Institute for Informatics, Faculty of Science, University of Amsterdam, 1098XH Amsterdam, The Netherlands

Abstract. *Optimization of binding affinities for ligands to their target protein is a primary objective in rational drug discovery. Herein we report on a collaborative study that evaluates various compounds designed to bind to the SET and MYND domain-containing protein 3 (SMYD3). SMYD3 is a histone methyltransferase and plays an important role in transcriptional regulation in cell proliferation, cell cycle and human carcinogenesis. Experimental measurements using the scintillation proximity assay show that the distributions of binding free energies from a large number of independent measurements exhibit non-normal properties. We use ESMACS (enhanced sampling of molecular dynamics with approximation of continuum solvent) and TIES (thermodynamic integration with enhanced sampling) protocols to rank the binding free energies and to provide detailed chemical insight into the nature of ligand−protein binding. Our results show that the 1-trajectory ESMACS protocol works well for the set of ligands studied here. Although one unexplained outlier exists, we obtain excellent statistical rankings across the set of compounds from the two protocols. ESMACS and TIES are again found to be powerful protocols for the accurate comparison of the binding free energies.*

1. Introduction

SMYD3 has been characterized as a versatile lysine methyltransferase, and is associated with multiple types of cancer, including colorectal, liver, and breast cancer. A range of histone and non-histone protein substrates are lysine N-methylated by methyl transfer from the SAM cofactor of SMYD3. Notably, loss of SMYD3 catalytic activity inhibited tumorigenesis in the presence of oncogenic Ras[1], suggesting that inhibition of SMYD3 in cancers with elevated RAS pathway signalling may be a useful therapeutic strategy. However, few SMYD3 inhibitors have been reported[2]. GSK has previously reported a crystal structure of the ternary complex of SMYD3, SAH and MEKK2 and a second crystal structure showing GSK2807 binding in the SAM pocket[3]. More recently, the identification and optimisation of a series of isoxazole amides as SMYD3 inhibitors was reported[4]. The ability of computational binding free energy calculations to predict the affinity of that series of ligands is presented here.

The last ten years have seen substantial progress in the use of computational chemistry methods within both academia and the pharmaceutical industry for quantitative structure-based drug discovery, thanks to burgeoning computational power, the increasing number of crystal structures, the accuracy of force fields, the improvement of the sampling methods and control of errors, alongside the automation and general usability of the approaches. Many

pharmaceutical companies have adopted free energy predictions as a routine tool to support their drug discovery efforts[5-6]. This progress has been prompted by Schrödinger's drug discovery platform, especially the FEP+ implementation[7]. There are other packages and workflows used in academia and/or industry, which integrate and automate the process of free energy calculation, including the steps of planning, set up, simulation, and analyses[8-9].

In the last few years, our team at UCL has developed two ensemble-based protocols for free energy calculations, termed "enhanced sampling of molecular dynamics with approximation of continuum solvent" (ESMACS)[8,10] and "thermodynamic integration with enhanced sampling" (TIES)[8,11]. ESMACS is based on the molecular mechanics Poisson-Boltzmann surface area method (MMPBSA)[12] while TIES centres on thermodynamic integration (TI). Although the protocols are built around the standard MMPBSA and TI methodologies, the names and abbreviations of these protocols are used to emphasise the central importance of the ensemble based nature of the protocols employed[8,10,13-16]. The term "ensemble" here refers to a set of individual (often called "replica") simulations conducted for the same physical system, starting from different initial conformations[13] (and possibly also with varying model parameters[14]). Advances in high-end computing capabilities offer the opportunity to run all of the replicas concurrently, ensuring the results can be delivered rapidly, exactly as has been done in climate and weather forecasting for the past twenty years. Ensemble approaches lead to increased reliability and reproducibility, with tighter control of standard uncertainty for nonlinear systems which are chaotic in nature[8,17-18]. ESMACS and TIES are performed using a binding affinity calculator (BAC)[19] which is a computational pipeline to automate the processes of building, running and marshalling the molecular dynamics simulations, as well as collecting and analysing data.

Depending on the usability, reliability, rapidity and throughput, these automated packages could find application at various stages of the drug discovery process across the wider pharmaceutical industry. In practise, however, the application of computational approaches is still dependent on the experience and knowledge of the practitioner. It remains a challenge for non-expert users to apply these existing tools to make robust predictions on a timescale that can substantially impact drug discovery programmes. For a given approach, the success of predictions also varies significantly across different protein targets with different sets of compounds. Studies have shown that the initial crystal structures and the existence of multiple conformations can have a significant effect on the quality of free energy predictions[20-21]. Based on the experience, knowledge and intuition we have accumulated, we propose the following criteria to predict the quality of the calculations: 1) how well the binding site is defined and structured; 2) how well a compound fits into the binding pocket; and 3) how many rotamers and/or binding poses a compound may manifest.

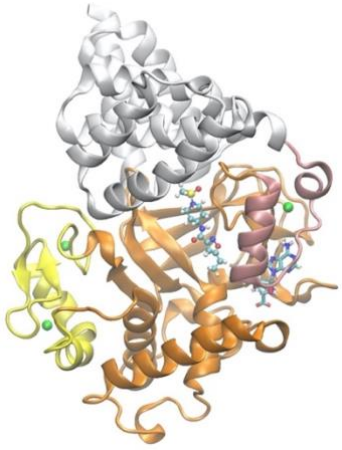


Figure 1. Structure of SMYD3, complexed with one of the compounds in this study. The N-terminal SET domain, the MYND domain, the post-SET domain and the C-terminal region are coloured in orange, yellow, pink and white, respectively. The cofactor SAH is shown with a bond representation, and the ligand (S01) in a ball-and-stick model. The Zn^{2+} ions are shown with sphere models and coloured green.

The purpose of the present study is to evaluate the ability of ESMACS and TIES to estimate binding affinities of a set of 22 ligands (Table 1) to the protein target. For the SMYD3 systems studied here, the binding site of the protein is well structured (Figure 1), and the binding mode for the scaffold of the congeneric compounds (Table 1) is well-defined in the crystal structure. It is thus likely, based on our foregoing criteria, that a reasonable prediction can be achieved, although the relatively large size of the binding site, the presence of multiple components in the structure, and the rotatable bonds at the R2 site of the compounds (Table 1) still pose a challenge for the conformational sampling and hence the accuracy and precision of the predictions.

2. Material and Methods

The x-ray structure of SMYD3 consists of a co-factor SAH (S-adenosylhomocysteine, a reaction product of the methyl group donor SAM (*S*-adenosyl-L-methionine)), and three zinc ions which are important for the folding of the protein (Figure 1).

2.1. Experiments

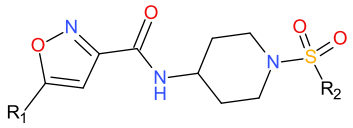
A series of isoxazole amides was chosen to cover a range of binding affinity and chemical structure. These include a diversity of lipophilicity, aromaticity and formal charge; the variation of formal charge has historically been challenging for free energy predictions. The IC50s acquired for the compounds were measured using the scintillation proximity assay (SPA) using MEKK2-based peptide as a substrate previously reported[4].

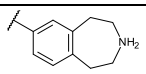
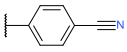
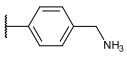
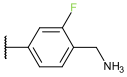
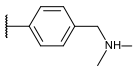
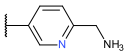
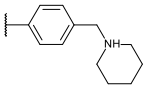
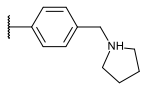
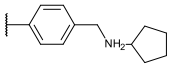
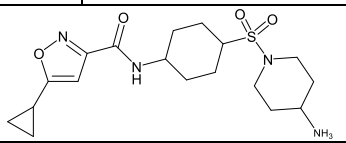
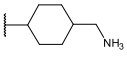
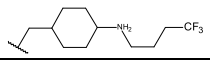
2.2. Computational approach

Compounds. A set of compounds named SXX were provided by GSK, where XX was a two-digit number where integers lower than ten were preceded by a 0 (Table 1), with mean values of pIC50 from experimental assay. Two of the compounds, S21 and S22, were reported as C01 and C28 in a previous publication[4]. All compounds shared the same scaffold (Table 1). The compounds were docked into the binding pocket of SMYD3 using Glide[22]. Modelling was carried out on a GSK internal structural precursor to 6P7Z. The rmsd between the structure used and 6P7Z is approximately 0.3 Å. The compounds were docked into the structure using glide XP with a substructural constraint on the isoxazole-amide-piperidine-sulfone substructure (as shown in table 1) using Glide in Maestro 2015-2.

Model preparation. Preparation and setup of the simulations were implemented using BAC[19], including parameterization of the compounds, solvation of the complexes, addition of counterions to electrostatically neutralize the systems and generation of configurations files for the simulations. The Amber package[23] was used for the parameterisation of the compounds, the set-up of the systems, and the analyses of the results. The Amber ff14SB force field was used for the protein, and TIP3P for water molecules. The protonation states were assigned using the "reduce" module of AmberTools. Parameters of the ligands were produced using the general Amber force field 2 (GAFF2) with Gaussian 16 to optimise compound geometries and to determine electrostatic potentials at the Hartree-Fock level with 6-31G** basis functions. The restrained electrostatic potential (RESP) module in the AmberTools was used to calculate the partial atomic charges for the compounds. All systems were solvated in orthorhombic water boxes with a minimum extension from the protein of 14 Å.

Table 1. Compounds investigated in this study.



Compound	R1	R2	pIC50	ΔG (kcal/mol)
S01	Et	Me	5.2	-7.14
S02	Et		5.5	-7.55
S03	Et	-CH(CH3)2	5.0	-6.86
S04	Et		5.5	-7.55
S05	Et		5.8	-7.96
S06	Et		5.1	-7.00
S07	Et		5.5	-7.55
S08			< 3.6[a]	> -4.94
S09			7.2	-9.88
S10		-CH2-CH2-NH3	6.4	-8.79
S11			5.4	-7.41
S12			7.0	-9.61
S13			7.1	-9.75
S14			6.9	-9.47
S15		-CH2-CH2-CH2-CH2-NH3	7.6	-10.43
S16			7.0	-9.61
S17			7.8	-10.71
S18			7.1	-9.75
S19			6.6	-9.06
S20			7.6	-10.43
S21		Me	5.3	-7.28
S22			8.5	-11.66

[a] There was no activity at the highest concentration (250 μM) tested.

ESMACS. We used the ESMACS (enhanced sampling of molecular dynamics with approximation of continuum solvent)[10] protocol for the simulations and analyses. The protocol uses replica simulations to obtain reproducible binding affinity predictions with robust uncertainty estimates. It is based on the molecular mechanics Poisson-Boltzmann surface area (MMPBSA), which is an approximate method for calculating absolute binding affinities from molecular dynamics trajectories. It is an endpoint free energy calculation, in which the difference in binding free energy, ΔG, is calculated using

$$\Delta G_{binding} = G_{com} - G_{pro} - G_{lig} \qquad Eq.\ 1$$

where G_i is the free energy of component *i* which corresponds to either complex (com), protein (pro), or ligand (lig), and is calculated from a set of structures from MD simulations. The free energies can be broken down into a number of components, including the molecular mechanics energy in the gas phase and the solvation free energy. While the former is derived from the molecular modelling forcefield used, the latter is estimated as the sum of the electrostatic solvation free energy calculated using the Poisson-Boltzmann equation and the nonpolar solvation free energy calculated from the solvent accessible surface area. The binding free energy is calculated from the difference between calculations performed for the complex, ligand and receptor conformations obtained from simulation. The 1-trajectory approach was used here, in which conformations of the protein and the ligands were extracted from the ligand-protein complex simulations.

TIES. We used thermodynamic integration with enhanced sampling (TIES)[11] to calculate the relative binding free energies for pairs of compounds. In TIES, an alchemical transformation for the mutated entity is used in both aqueous solution and within the ligand-protein complex. The free energy changes of the alchemical mutation processes, ΔG_{aq}^{alch} and $\Delta G_{complex}^{alch}$, are calculated by:

$$\Delta G^{alch} = \int_0^1 \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \qquad Eq.\ 2$$

Here λ ($0 \leq \lambda \leq 1$) is an alchemical coupling parameter such that λ=0 and λ=1 correspond to the initial and final thermodynamic states, and $\partial V(\lambda)/\partial \lambda$ is the partial derivative of the hybrid potential energy $V(\lambda)$ at an intermediate state λ. $\langle \cdots \rangle_\lambda$ denotes an ensemble average over configurations representative of the state λ.

The binding free energy difference is then calculated from

$$\Delta\Delta G^{binding} = \Delta G_{complex}^{alch} - \Delta G_{aq}^{alch} \qquad Eq.\ 3$$

Simulations. The binding affinity calculator (BAC)[19], an automated workflow tool for free energy calculations, was used to prepare and execute ESMACS and TIES simulations. The standard protocol[10-11,24] was used, in which NAMD[25] simulations with 25 and 5 replicas were performed for ESMACS and TIES, respectively. Each replica in the ensemble started with identical atomic coordinates, with different initial velocities generated independently from a Maxwell−Boltzmann distribution.

To avoid the well-known "end-point catastrophe"[26], a soft-core potential was used for van der Waals (vdW) interactions involving the perturbed atoms. The electrostatic interactions were linearly scaled but at a faster rate than the vdW interactions, so that the partial charges were removed for the disappearing atoms before they were fully annihilated, and were introduced on the appearing atoms after they already partially appeared.

The MD package NAMD2.12[25] was used throughout the equilibration and production runs of all simulations. For each replica in an ensemble, energy minimizations were first performed with heavy protein atoms restrained at their initial positions. The initial velocities were then generated independently from a Maxwell−Boltzmann distribution at 50 K, and the systems were heated up to and kept at 300K within 60 ps. A series of equilibration runs, totalling 2 ns, were conducted, while the restraints on heavy atoms were gradually reduced. Finally, 4 ns production simulations were run for each replica for all ESMACS and TIES simulations.

The ESMACS simulations for the compounds S01 – S20 were initially conducted using 10-replica ensembles on the DNAnexus platform (https://www.dnanexus.com/) which provides strong cybersecurity. Previous studies[10,24,27-29] have stablished a standard ESMACS protocol with 25 replicas, and shown that the combination of the simulation length and the size of the ensemble provides a trade-off between computational cost and precision. The choice of a smaller number here was designed to reduce computational cost on the cloud environment. The study was later extended to include two more compounds, C01 and C28 from Su *et al.*[4], renamed as S21 and S22, to extend ESMACS to 25 replicas and, more importantly, to perform TIES studies on the selected compound pairs. The Blue Waters supercomputer at the National Center for Supercomputing Applications (NCSA) in the US was used for the extended ESMACS simulations. The SuperMUC supercomputer at Leibniz Supercomputing Centre in Germany was used for the TIES simulations.

3. Results

To assess the accuracy and precision of the method, we evaluated the binding affinities of the ligands (Table 1) to SMYD3, and compared the computed results with experimental data. ESMACS was used for the full set of the ligands, while TIES was applied to some selected pairs of the ligands with the same net charge.

3.1. Reproducibility of the ESMACS simulations

It is well-studied that the differences of the initial conditions among individual simulations lead to rapid divergence of trajectories[13]. Many complex systems hence exhibit sensitive dependence on initial conditions. The calculated thermodynamic properties from individual simulations will therefore inevitably differ. Two sets of ESMACS simulations were performed for the complexes SXX (Table 1) independently on Blue Waters and DNAnexus (see the Material and Methods section above). Figure 2 shows the variances and correlation of the calculated binding free energies from the two sets of simulations. Excellent agreement was observed between calculations using two different computational platforms: HPC machine Blue Waters and cloud environment DNAnexus, with a highly significant Spearman correlation of 0.98. No statistical differences were seen between the two sets of calculated binding free energies: 16 out of 20 compounds having identical results, within error bars, and the remaining 4 within two error bars (Figure 2). The two simulations produce consistent results, with a mean signed difference of 0.13 kcal/mol and a mean unsigned difference of 0.63 kcal/mol. Both of the calculations have good correlations with the experimental measurement, with correlation coefficients of 0.80 and 0.78 for the simulations on Blue Waters and DNAnexus, respectively. Because of the smaller number of replicas used in the DNAnexus simulations, the error bars in these simulations are ~1.5 times larger than those from Blue Waters simulations. As the two simulations produce similar accuracies, only the results from Blue Waters are reported in the following analyses.

3.2. Correlations between ESMACS calculations and experimental measurements

The predicted binding free energies from the 1-trajectory approach exhibit a high correlation with the experimental data (Figure 3), with a Pearson correlation coefficient of 0.84. In a pharmaceutical drug development project, compounds are designed or selected for the same protein target. The ranking of the binding affinities is not affected by the energy of the protein G_{pro} (Eq. 1) when the conformational space is sufficiently sampled. Free energies of a protein differ in its bound and unbound states. The difference, called the adaptation free energy[10], provides an indication of the conformational changes of the protein and the energetic costs upon binding. Inclusion of adaptation free energies improves the correlations between simulations and experimental measurements in some cases[10,30-31], and does not have obvious effects in other cases[32]. For the current data set of compounds, the binding site is relatively large. No significant strain is induced within protein upon compound binding. The inclusion of adaptation free energies of the protein degrades the correlations, with a correlation coefficient of 0.70.

The calculations correctly distinguish the charged compounds from the neutral ones (Figure 3). The R2 group (Table 1) locates in a hydrophilic pocket in which negatively charged residues GLU192, ASP241 and GLU294 form favourable electrostatic interactions with the positively charged R2 group (Figure 4). This makes the binding of charged compounds more favourable in general than the electrostatically neutral ones. The two variants at R1 studied here may not affect the binding affinities significantly because the ethyl group and the 3-membered ring are similar in their hydrophobic properties and their sizes. The two compounds, S01 and S21, differing only at the R1 group, have similar binding affinities from both experimental measurements and the ESMACS calculations (Figure 3). It should be noted that no activity was detected for S08 at the highest concentration (250 μM) in the experiments (Table 1), indicating that its binding affinity is likely to be less negative than that presented in Figure 3. This makes the point deviating even farther from the regression line. Our TIES calculations also show that S08 is an outlier (see details below).

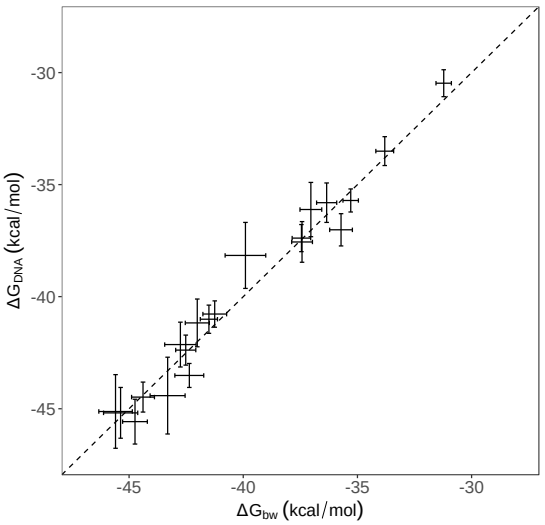


Figure 2. Comparison of calculated binding free energies from two independent studies of the ligand-SMYD3 models performed on Blue Waters (bw, horizontal axis) and DNAnexus (DNA, vertical axis). Dashed line shows an ideal y=x regression. The standard errors are calculated using a bootstrapping method.

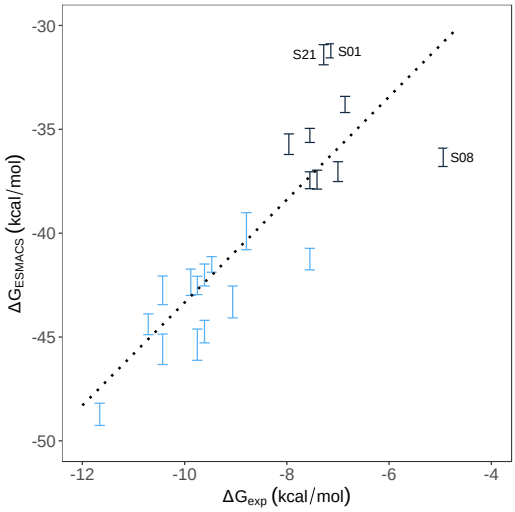


Figure 3. Comparison of calculated binding free energies and experimental binding affinity data from 1-traj ESMACS approach. The dotted line shows a linear regression. A correlation coefficient of 0.84 is obtained for the entire set of compounds. The +1e changed compounds are shown in blue, and electrostatically neutral ones in black.

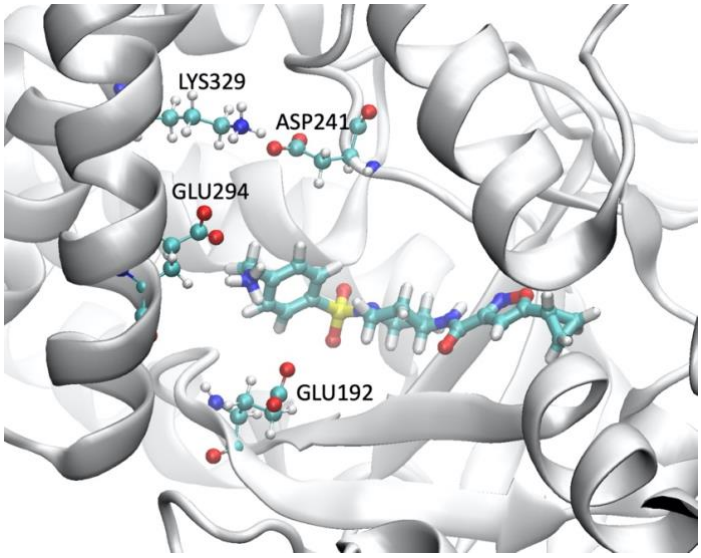


Figure 4. Electrostatic interactions between the R2 group and the protein. The negatively charged residues GLU192, ASP241, GLU294, and positively charged R2 group, along with positively charged LYS 329, form favourable electrostatic interactions.

3.3. TIES results

Relative binding free energies (ΔΔG) are calculated using TIES for selected pairs of the compounds. Each compound is paired at least once with other compounds. No compounds are paired if they have different net charges, as alchemical methods encounter specific difficulties when changes in the net charge arise and charge corrections are required. The results of these relative binding free energy calculations are compared with the data derived from experimental measurements (Figure 5).

As the compound S08 may be denoted as an outlier (see details below), the analyses are performed separately for the dataset with and without the compound. The overall mean unsigned error (MUE) is 1.21 kcal/mol for the entire dataset, and 0.68 kcal/mol when pairs involving S08 are excluded. The mean signed errors (MSEs) are 0.62 kcal/mol and -0.06 kcal/mol for the dataset with and without S08, respectively. Except the pairs with S08, only one compound pair, S22–S20, has a predicted ΔΔG value which differs from the experimental data by >2 kcal/mol. The main difference between the two compounds are the 3 rotatable bonds at R2 (Table 1). The rotation of these rotatable bonds leads to large conformational flexibilities in S22, which may need longer simulation time to get reliable prediction.

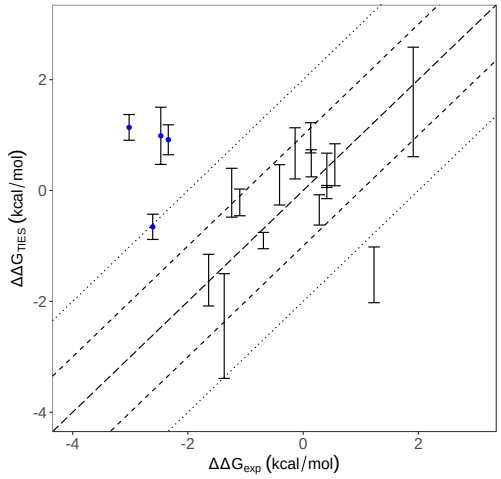


Figure 5. Correlation between TIES-predicted relative binding affinities and experimental data for a total of 19 compound pairs. The long dashed line represents y=x, whereas dashed and dotted lines represent 1 kcal/mol and 2 kcal/mol ranges, respectively. The pairs involving S08 are highlighted in blue.

3.4. S08 remains an outlier

Our initial TIES calculations only contained one compound pair involving S08, of which the deviation between the calculation and the experimental data is large. As the ESMACS simulation also shows it as an outlier, we have paired S08 with 3 other compounds for TIES simulations. The results only confirm that there is a systematic deviation in the binding free energy for S08 between calculations and experimental measurements. Based on the 4 TIES calculations for S08, the average difference between calculations and experimental data is 3.21 kcal/mol. In another word, the compound S08 needs to have a binding affinity 3.21 kcal/mol more negative in the experiments, or 3.21 kcal/mol less negative in the calculations, to make them agree with each other. This value is also in a good agreement with the ESMACS calculation, with which the data point for S08 can be shifted much closer to the regression line (Figure 3).

The compound S08 consists of a nitrile group of which the nitrogen is highly electronegative. Although the negatively charged residues at the R2 pocket are unfavourable for the presentence of the nitrile group, the positively charged residue LYS329 and the relatively spacious pocket appear to be able to tolerate the group. The detailed analyses of the simulation trajectories do not provide more insights. Further searching in the experimental data set has identified another compound which is very similar to S08 and shares the same nitrile group at the R2. The compound also does not show any activities at the highest concentration tested in the assay (data not shown). Although it could be an experimental issue, it is more likely to be a force field or possibly sampling issue. As there are no satisfactory explanations for the disagreement between the experiments and the calculations, S08 remains as an unexplained outlier. Such unexplained outliers are not unusual in drug discovery and development projects. Machine-learning approaches have been proposed to identify the differences between the calculations and the experimental data, and to provide empirical correction terms to the predictions from the alchemical approaches but these too depend on assumptions which are rarely articulated concerning the way in which the data are distributed[33].

3.5. Non-normal distributions of free energy calculations and measurement

Normal distributions have been typically assumed in experimental measurements and calculations of binding free energies. The assumptions are commonly made for the true $\Delta G_{binding}$ for a large number of compounds, for the experimentally determined and computationally predicted $\Delta G_{binding}$ for a given compound, as well as for the relative binding free energies $\Delta\Delta G_{binding}$. The normal distributions are characterised by an average μ and a standard deviation σ. Although the presence of uncertainties is known to the scientific community broadly, they are still "known unknowns": in many cases we do not know the order of magnitude of the various uncertainties, the sources and the consequences of them, not to mention how to reduce them. It is important to describe the free energy distributions carefully as many statistical analyses are based on it. The most important assumption in regression dilution[34], for example, is that all the variables under consideration are normally distributed. If this is not the case, regression dilution may not be applied.

The assertion that the calculated binding free energies ΔG_{cal} or binding free energy differences $\Delta\Delta G_{cal}$ follow a normal distribution conflicts with our observation that such data are not in general normally distributed[8,16,18,29]. Newtonian molecular dynamics is inherently nonlinear, and this is the underlying reason why the dynamics is chaotic in the technical sense. Not only are individual trajectories extremely sensitive to initial conditions, they become increasingly

inaccurate as the duration of such a simulation unfolds. They manifest long range correlations which are not present in Gaussian statistics.

In experimental measurement of binding free energies, uncertainties on the order of 0.3 − 0.5 kcal/mol for ΔG_{exp} and 0.4 – 0.7 kcal/mol for $\Delta\Delta G_{exp}$ have been claimed from high-quality experimental measurements[7]. It is, however, very often the case that experimental data are reported as single numbers, without quantification of the uncertainties. We have no knowledge about the statistics of ΔG_{exp} or $\Delta\Delta G_{exp}$ reported, let alone the distribution of these quantities. This means that the unknown and unstated error bars may be varying in all manner of ways, so claiming that they are normally distributed is not credible.

There are four compounds in the current project, which have been tested >100 times for their activities to SMYD3. One of them is S21 (Figure 6a) which has been computationally investigated here. The other three are for more potent compounds that are from a related but slightly different series. The relatively large number of tests makes it possible to verify the distributions of the experimental data. It should be noted that while compounds a and b do not show any drift in the assay over time, compounds c and d (Table 2 and Figure 6) show a small amount of time dependency.

All of the 4 distributions are skewed from a normal distribution, with skewness deviating from 0. Three of them are highly skewed with positive skewness (Table 2), indicating that the distributions have longer tails on the right side than those on the left (Figure 6). The other one is moderately skewed with a negative skewness and a longer tail on the left. The excess kurtoses are all positive, meaning that compared to a normal distribution, the tails are longer and heavier. It should be noted that the experimental data shown in Figure 6 are representative of the behaviour of such ligand-protein binding affinity more widely, and that other data remains confidential to GSK. Overall, these results imply the presence of non-normal distributions in the experimental measurements.

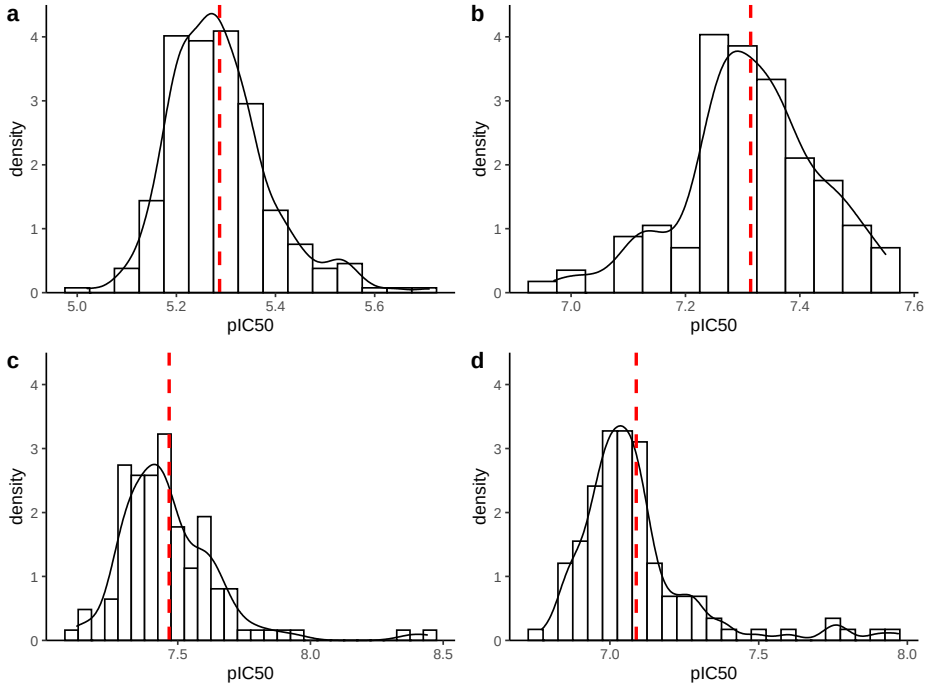


Figure 6. Distributions of experimental pIC50 values shown as histogram and kernel density curve. 4 compounds from the current project (a-d) are used, which have been tested more than 100 times. A bin size of .05 is used. The dashed lines indicate the means of the experimental measurements.

Table 2. Statistics of the experimental measurements pIC50 for the compounds with >100 tests.

compound	No. of test	Average	sd	skewness	kurtosis
a	264	5.29	0.10	0.88	1.47
b	114	7.31	0.12	-0.35	0.23
c	124	7.47	0.19	2.04	7.56
d	116	7.09	0.21	2.11	5.30

4. Conclusion

Using the TIES and ESMACS protocols, we have computed the binding free energies of a series of ligands to zinc finger protein SMYD3. Although an unexplained outlier exists, we obtain excellent statistical rankings across the set of compounds from the two protocols. ESMACS and TIES are again found to be powerful protocols for the accurate comparison of the binding free energies.

We have previously reported the non-normal properties of calculated binding free energies. In the current study, we investigate the distributions of experimentally measured free energies, and find that the distributions are highly skewed. The practical implications of this discovery are important to apprehend. Non-normal distributions imply the occurrence of more 'outliers', making it essential to perform multiple measurements to pin down average behaviour. It is also a call to exercise caution in the use of statistical methods for the comparison of experimental data and computational predictions, as the assumption of normal distributions is not generally valid.

Acknowledgments

We are grateful for funding for the UK MRC Medical Bioinformatics project (grant no. MR/L016311/1), the EPSRC funded UK Consortium on Mesoscale Engineering Sciences (UKCOMES grant no. EP/L00030X/1), the European Commission for EU H2020 CompBioMed2 Centre of Excellence (grant no. 823712) and for EU H2020 EXDCI-2 project (grant no. 800957), and NSF Award (https://www.nsf.gov/pubs/2017/nsf17542/nsf17542.htm, Award No. NSF 1713749). We made use of the Blue Waters supercomputer at the National Center for Supercomputing Applications of the University of Illinois at Urbana-Champaign, access to which was made available through the aforementioned NSF award. Additional calculations were conducted using an award of computer time on the Theta machine at Argonne Leadership Computing Facility provided by the US Department of Energy's Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program (through the INSPIRE project and a 2021 award "COMPBIO"). We are grateful to DNAnexus for both computational resources and technical support from Fiona Ford, Brett Hannigan and Chai Fungtammasan. We acknowledge the Leibniz Supercomputing Centre for providing access to SuperMUC (https://www.lrz.de/services/compute/) and the very able assistance of its scientific support staff.

Author Information

Corresponding Author

Peter V. Coveney − Centre for Computational Science, Department of Chemistry, University College London, London WC1H0AJ, United Kingdom; Advanced Research Computing Centre, Department of Chemistry, University College London, London WC1H 0AJ, United

Kingdom; Informatics Institute, University of Amsterdam, Amsterdam 1012 WX, The Netherlands; orcid.org/0000-0002-8787-7256; Email: p.v.coveney@ ucl.ac.uk

Authors

Shunzhou Wan − Centre for Computational Science, Department of Chemistry, University College London, London WC1H0AJ, United Kingdom; orcid.org/0000-0001-7192-1999

Agastya P. Bhati − Centre for Computational Science, Department of Chemistry, University College London, London WC1H0AJ, United Kingdom; orcid.org/0000-0003-4539-4819

David W. Wright − Centre for Computational Science, Department of Chemistry, University College London, London WC1H0AJ, United Kingdom; orcid.org/0000-0002-5124-8044

Ian Wall − GlaxoSmithKline, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, United Kingdom.

Alan Graves − GlaxoSmithKline, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, United Kingdom.

Darren Green − GlaxoSmithKline, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, United Kingdom.

References

1. Mazur, P. K.; Reynoird, N.; Khatri, P.; Jansen, P. W.; Wilkinson, A. W.; Liu, S.; Barbash, O.; Van Aller, G. S.; Huddleston, M.; Dhanak, D.; Tummino, P. J.; Kruger, R. G.; Garcia, B. A.; Butte, A. J.; Vermeulen, M.; Sage, J.; Gozani, O., Smyd3 Links Lysine Methylation of Map3k2 to Ras-Driven Cancer. *Nature* **2014,** *510*, 283-287.
2. Bottino, C.; Peserico, A.; Simone, C.; Caretti, G., Smyd3: An Oncogenic Driver Targeting Epigenetic Regulation and Signaling Pathways. *Cancers (Basel)* **2020,** *12*.
3. Van Aller, G. S.; Graves, A. P.; Elkins, P. A.; Bonnette, W. G.; McDevitt, P. J.; Zappacosta, F.; Annan, R. S.; Dean, T. W.; Su, D. S.; Carpenter, C. L.; Mohammad, H. P.; Kruger, R. G., Structure-Based Design of a Novel Smyd3 Inhibitor That Bridges the Sam-and Mekk2-Binding Pockets. *Structure* **2016,** *24*, 774-781.
4. Su, D. S.; Qu, J.; Schulz, M.; Blackledge, C. W.; Yu, H.; Zeng, J.; Burgess, J.; Reif, A.; Stern, M.; Nagarajan, R.; Pappalardi, M. B.; Wong, K.; Graves, A. P.; Bonnette, W.; Wang, L.; Elkins, P.; Knapp-Reed, B.; Carson, J. D.; McHugh, C.; Mohammad, H.; Kruger, R.; Luengo, J.; Heerding, D. A.; Creasy, C. L., Discovery of Isoxazole Amides as Potent and Selective Smyd3 Inhibitors. *ACS Med Chem Lett* **2020,** *11*, 133-140.
5. Schindler, C. E. M.; Baumann, H.; Blum, A.; Bose, D.; Buchstaller, H. P.; Burgdorf, L.; Cappel, D.; Chekler, E.; Czodrowski, P.; Dorsch, D.; Eguida, M. K. I.; Follows, B.; Fuchss, T.; Gradler, U.; Gunera, J.; Johnson, T.; Jorand Lebrun, C.; Karra, S.; Klein, M.; Knehans, T.; Koetzner, L.; Krier, M.; Leiendecker, M.; Leuthner, B.; Li, L.; Mochalkin, I.; Musil, D.; Neagu, C.; Rippmann, F.; Schiemann, K.; Schulz, R.; Steinbrecher, T.; Tanzer, E. M.; Unzue Lopez, A.; Viacava Follis, A.; Wegener, A.; Kuhn, D., Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. *J Chem Inf Model* **2020,** *60*, 5457-5474.

6. Ciordia, M.; Pérez-Benito, L.; Delgado, F.; Trabanco, A. A.; Tresadern, G., Application of Free Energy Perturbation for the Design of Bace1 Inhibitors. *J Chem Inf Model* **2016,** *56*, 1856-1871.
7. 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, R.; 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.
8. Wan, S.; Bhati, A. P.; Zasada, S. J.; Coveney, P. V., Rapid, Accurate, Precise and Reproducible Ligand-Protein Binding Free Energy Prediction. *Interface Focus* **2020,** *10*, 20200007.
9. Wan, S.; Tresadern, G.; Pérez-Benito, L.; Vlijmen, H.; Coveney, P. V., Accuracy and Precision of Alchemical Relative Free Energy Predictions with and without Replica-Exchange. *Adv Theory Simul* **2019,** *3*, 1900195.
10. Wan, S.; Knapp, B.; Wright, D. W.; Deane, C. M.; Coveney, P. V., Rapid, Precise, and Reproducible Prediction of Peptide-MHC Binding Affinities from Molecular Dynamics That Correlate Well with Experiment. *J Chem Theory Comput* **2015,** *11*, 3346-3356.
11. Bhati, A. P.; Wan, S.; Wright, D. W.; Coveney, P. V., Rapid, Accurate, Precise, and Reliable Relative Free Energy Prediction Using Ensemble Based Thermodynamic Integration. *J Chem Theory Comput* **2017,** *13*, 210-222.
12. Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham III, T. E., Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Acc Chem Res* **2000,** *33*, 889-897.
13. Coveney, P. V.; Wan, S., On the Calculation of Equilibrium Thermodynamic Properties from Molecular Dynamics. *Phys Chem Chem Phys* **2016,** *18*, 30236-30240.
14. Vassaux, M.; Wan, S.; Edeling, W.; Coveney, P. V., Ensembles Are Required to Handle Aleatoric and Parametric Uncertainty in Molecular Dynamics Simulation. *J Chem Theory Comput* **2021,** *17*, 5187-5197.
15. Wade, A.; Bhati, A. P.; Wan, S.; Coveney, P. V., Alchemical Free Energy Estimators and Molecular Dynamics Engines: Accuracy, Precision and Reproducibility. *ChemRxiv* **2021**, DOI: 10.26434/chemrxiv-22021-nqp26438r.
16. Bhati, A. P.; Coveney, P. V., Large Scale Study of Ligand-Protein Relative Binding Free Energy Calculations: Actionable Predictions from Statistically Robust Protocols. *ChemRxiv* **2021**, DOI: 10.26434/chemrxiv-22021-zdzng.
17. 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.
18. Wan, S.; Sinclair, R. C.; Coveney, P. V., Uncertainty Quantification in Classical Molecular Dynamics. *Philos Trans R Soc A* **2021,** *379*, 20200082.
19. Sadiq, S. K.; Wright, D.; Watson, S. J.; Zasada, S. J.; Stoica, I.; Coveney, P. V., Automated Molecular Simulation Based Binding Affinity Calculator for Ligand-Bound HIV-1 Proteases. *J Chem Inf Model* **2008,** *48*, 1909-1919.
20. Suruzhon, M.; Bodnarchuk, M. S.; Ciancetta, A.; Viner, R.; Wall, I. D.; Essex, J. W., Sensitivity of Binding Free Energy Calculations to Initial Protein Crystal Structure. *J Chem Theory Comput* **2021,** *17*, 1806-1821.

21. Pérez-Benito, L.; Keränen, H.; van Vlijmen, H.; Tresadern, G., Predicting Binding Free Energies of Pde2 Inhibitors. The Difficulties of Protein Conformation. *Sci Rep-Uk* **2018,** *8*, 4883.
22. Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T., Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J Med Chem* **2006,** *49*, 6177-6196.
23. Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber Biomolecular Simulation Programs. *J Comput Chem* **2005,** *26*, 1668-1688.
24. Wright, D. W.; Hall, B. A.; Kenway, O. A.; Jha, S.; Coveney, P. V., Computing Clinically Relevant Binding Free Energies of HIV-1 Protease Inhibitors. *J Chem Theory Comput* **2014,** *10*, 1228-1241.
25. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K., Scalable Molecular Dynamics with NAMD. *J Comput Chem* **2005,** *26*, 1781-1802.
26. Beveridge, D. L.; Dicapua, F. M., Free-Energy Via Molecular Simulation - Applications to Chemical and Biomolecular Systems. *Annu Rev Biophys Bio* **1989,** *18*, 431-492.
27. Genheden, S.; Ryde, U., How to Obtain Statistically Converged MM/GBSA Results. *J Comput Chem* **2010,** *31*, 837-846.
28. Sadiq, S. K.; Wright, D. W.; Kenway, O. A.; Coveney, P. V., Accurate Ensemble Molecular Dynamics Binding Free Energy Ranking of Multidrug-Resistant HIV-1 Proteases. *J Chem Inf Model* **2010,** *50*, 890-905.
29. Bieniek, M. K.; Bhati, A. P.; Wan, S.; Coveney, P. V., Ties 20: Relative Binding Free Energy with a Flexible Superimposition Algorithm and Partial Ring Morphing. *J Chem Theory Comput* **2021,** *17*, 1250-1265.
30. Wan, S.; Bhati, A. P.; Skerratt, S.; Omoto, K.; Shanmugasundaram, V.; Bagal, S. K.; Coveney, P. V., Evaluation and Characterization of Trk Kinase Inhibitors for the Treatment of Pain: Reliable Binding Affinity Predictions from Theory and Computation. *J Chem Inf Model* **2017,** *57*, 897-909.
31. Wan, S.; Bhati, A. P.; Zasada, S. J.; Wall, I.; Green, D.; Bamborough, P.; Coveney, P. V., Rapid and Reliable Binding Affinity Prediction of Bromodomain Inhibitors: A Computational Study. *J Chem Theory Comput* **2017,** *13*, 784-795.
32. Wright, D. W.; Husseini, F.; Wan, S.; Meyer, C.; van Vlijmen, H.; Tresadern, G.; Coveney, P. V., Application of the ESMACS Binding Free Energy Protocol to a Multi‐Binding Site Lactate Dehydogenase a Ligand Dataset. *Adv Theory Simul* **2019,** *3*, 1900194.
33. Scheen, J.; Wu, W.; Mey, A.; Tosco, P.; Mackey, M.; Michel, J., Hybrid Alchemical Free Energy/Machine-Learning Methodology for the Computation of Hydration Free Energies. *J Chem Inf Model* **2020,** *60*, 5331-5339.
34. Frost, C.; Thompson, S. G., Correcting for Regression Dilution Bias: Comparison of Methods for a Single Predictor Variable. *J R Stat Soc* **2000,** *163*, 173-189.