26 September 2022

Non-Gaussian distributions of absolute free energies in ensemble molecular dynamics simulations

Peter V Coveney[1,2,3*], *Shunzhou Wan*[1]

[1]Centre for Computational Science, Department of Chemistry, University College London, London WC1H 0AJ, U. K.
[2]Advanced Research Computing Centre, University College London, London WC1H 0AJ, U.K.
[3]Institute for Informatics, Faculty of Science, University of Amsterdam, 1098XH Amsterdam, The Netherlands
*Corresponding to: p.v.coveney@ucl.ac.uk

Abstract:

Significantly more 'outliers' can be produced from a non-Gaussian distribution than one would anticipate were the statistics to conform to a normal distribution. Using ensemble simulations consisting of 25 replicas, we have previously identified a considerable percentage of ligand-protein systems which present non-Gaussian distributions in calculated binding free energies. Here we report on the statistics of much larger ensembles and find that the free energy distributions are definitively non-Gaussian for these systems.

Introduction

Non-Gaussian statistics has been reported for binding free energies calculated from molecular dynamics (MD)[1, 2] and Monte Carlo (MC)[3] simulations, as well as in experiments such as the scintillation proximity assay[4]. Non-Gaussian distributions have also been observed for geometrical quantities emanating from large ensembles of MD simulations of a T-cell receptor/peptide/major histocompatibility complex[5]. Such non-Gaussian properties cannot be unambiguously observed without large number of samples. When the underlying distribution deviates from normal, it is likely to exhibit a significantly higher frequency of occurrence of so-called "rare events". This makes it harder to infer the real value for a property of interest from single observations.

We have recently applied the standard ESMACS (ensembled-based enhanced sampling of molecular dynamics with approximation of continuum solvent) protocol[1] to investigate binding free energy distributions for approximately 400 ligand–protein complexes[2], in which an ensemble of 25 replicas is used. The Shapiro-Wilk and D'Agostino/Pearson normality tests show that 27% and 20% respectively of the complexes reject the null hypothesis of normal distribution, at the level of significant 0.01. In other words, a significant percentage of these molecular systems exhibit deviations from the standard Gaussian profile for the frequency distribution of predicted binding free energies from ensemble MD simulations composing 25 replicas. Although this number of replicas is already deemed large and not routinely used by most bio-MD practitioner other than ourselves in free energy calculations, it is still not big enough to be able to quantify these statistics definitively. The calculated skewness and kurtosis, for example, have sizeable uncertainties (Table 1). Here we select a subset of the aforementioned ESMACS dataset, and increase the number of replicas by more than one order of magnitude. The large number of replicas provides us with a definitive view about the distributions of the predicted binding free energies.

Methods

To obtain definitive conclusions concerning the distribution of predicted absolute binding free energies, we have applied an extended ESMACS study with a large ensemble consisting of 500 replicas. Nine ligand–protein complexes are selected from our previous study[2], which have the largest skewness and kurtosis. The study predicts binding free energies for a large set of compounds targeting SARS-COVID-2 proteins. We use the models we prepared previously, in which Amberff14SB[6] is used for the proteins and GAFF2 for the compounds. The AmberTools package[7] was used for the parameterization of the compounds and the preparation of the systems[1], and is used here for the analyses of the results for the extended ensembles. All simulations were run on Summit at the Oak Ridge National Laboratory using NAMD3[8].

Results

The distributions of the predicted binding free energies are summarised graphically in Figure 1. The probability plots show clearly one or more of the followings: 1) the differences between the means and the modes, 2) the skewness, 3) the kurtosis, 4) the long and heavy tail(s), and 5) the presence of multiple modes in the predicted binding free energies. The convergence of excess skewness and kurtosis with the number of replicas is also definitive, showing the two quantities unambiguously deviating from 0 in ensemble simulations with a sufficiently large number of replicas (Figure 2). Although the skewness and kurtosis are not decisive for some molecular systems within 25-replica simulations, they are definitely non-zero from 500-replica simulations (Table 1 and Figure 2). The same is true for the Shapiro-Wilk and D'Agostino/Pearson normality tests, demonstrating that although 25-replica results are not definitive for some molecular systems, the tests unequivocally reject the normal null hypothesis with very high confidence when the number of replicas is increased to 500 (Table 2).

Table 1. Skewness and excess kurtosis of the calculated binding free energy distributions. Errors are given in brackets, calculated at 95% confidence interval using bootstrapping.

complex	25-replica ensemble		500-replica ensemble	
	Skewness	Kurtosis	Skewness	Kurtosis
a	-2.58 [-4.51, -1.67]	7.21 [2.12, 15.30]	-0.84 [-1.11, -0.57]	1.36 [0.50, 2.14]
b	-0.31 [-0.91, 0.27]	-0.80 [-1.90, -0.23]	-0.43 [-0.57, -0.29]	-1.18 [-1.39, -1.03]
c	-0.15 [-0.79, 0.55]	-1.44 [-2.44, -1.16]	-0.15 [-0.30, 0.00]	-1.18 [-1.36, -1.03]
d	-2.60 [-4.98, -1.78]	7.71 [2.96, 16.08]	-0.87 [-1.72, -0.24]	4.84 [2.28, 8.47]
e	-2.87 [-5.61, -2.14]	9.29 [5.10, 19.37]	-1.73 [-1.93, -1.50]	2.57 [1.39, 3.51]
f	0.35 [-0.87, 1.30]	0.51 [-1.45, 2.18]	1.27 [1.03, 1.50]	2.03 [1.15, 2.78]
g	2.07 [1.34, 4.05]	5.22 [1.50, 11.62]	1.07 [0.68, 1.53]	3.08 [1.70, 4.55]
h	0.91 [-0.13, 1.63]	-0.75 [-5.22, 0.27]	0.44 [0.26, 0.63]	-0.14 [-0.59, 0.28]
i	0.26 [-0.35, 0.87]	-1.17 [-2.24, -0.77]	-0.10 [-0.24, 0.03]	-0.69 [-0.89, -0.52]

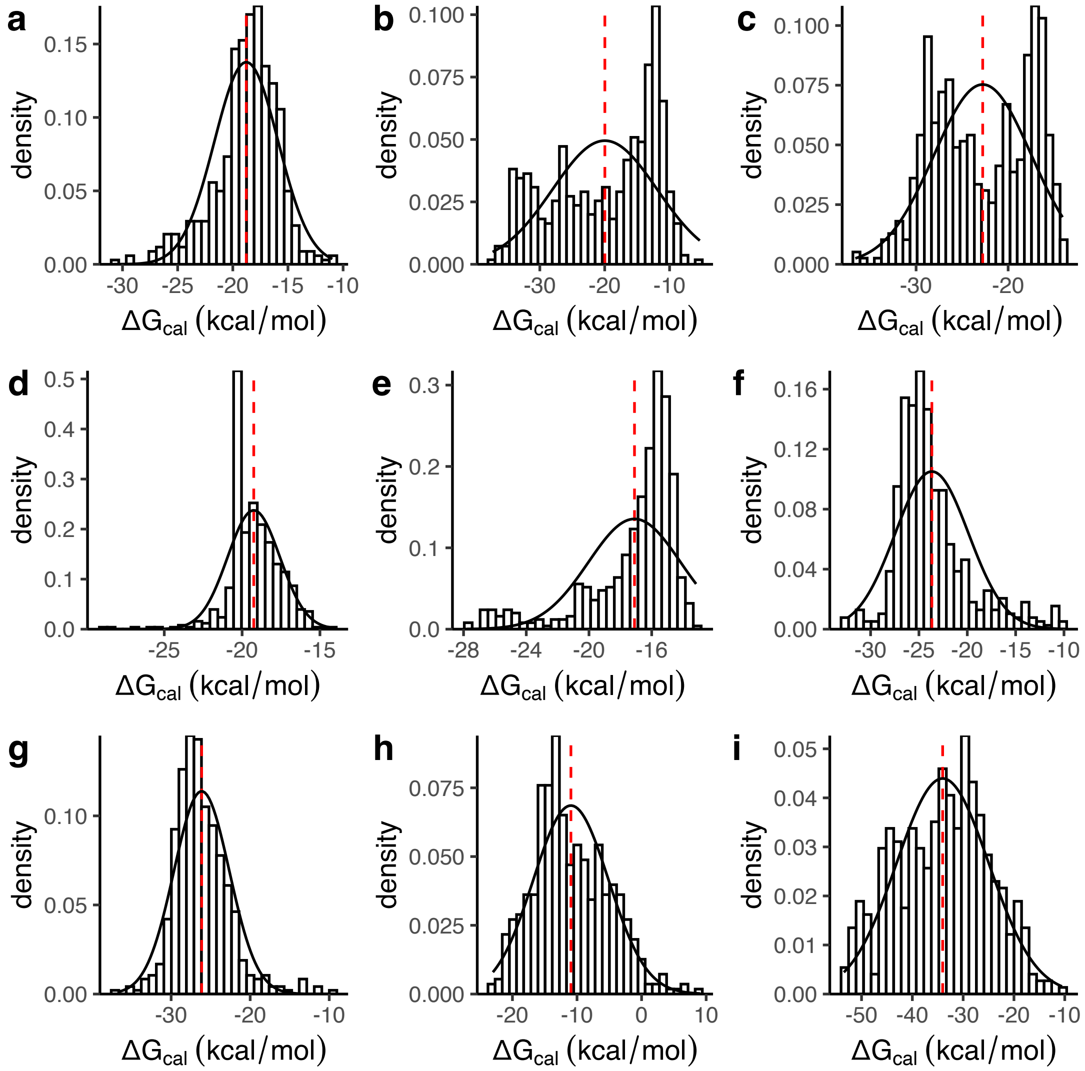

Figure 1. Distributions of predicted absolute binding free energies (ΔG) using the ESMACS approach with large ensembles. Nine ligand-protein complexes have been investigated by ESMACS with 500 replicas each. The best-fit Gaussian distributions are shown by black solid lines, while the red dashed lines indicate the average values.

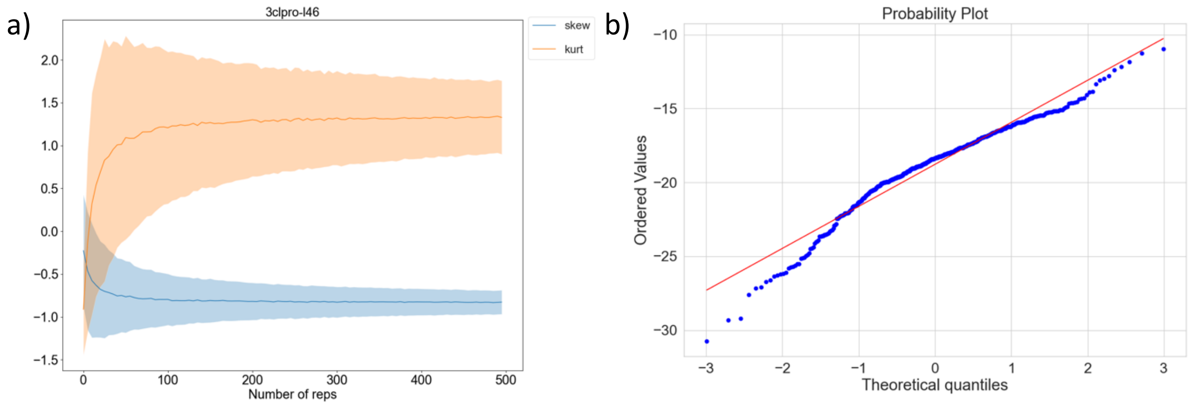

Figure 2. The skewness and kurtosis of the binding free energy distribution for one ligand-protein complex investigated (subfigure a in Figure 1). The convergence of the skewness and kurtosis a), with means (solid lines) and standard errors of the mean (shaded region). The quantile-quantile plot b) shows the deviation of the quantiles (blue dots) from an ideal Q–Q plot (red line), clearly exhibiting the non-normal distributions from calculations.

Table 2. Confidence (p-value) that the null hypothesis is false from Shapiro-Wilk and D'Agostino/Pearson normality tests.

complex	*p*-value (25 reps)		*p*-value (500 reps)	
	Shapiro-Wilk	Pearson	Shapiro-Wilk	Pearson
a	5.22×10^{-6}	5.03×10^{-8}	5.58×10^{-11}	1.60×10^{-14}
b	0.524	0.537	3.46×10^{-16}	5.59×10^{-64}
c	0.032	0.008	6.03×10^{-13}	1.17×10^{-60}
d	1.46×10^{-5}	3.24×10^{-8}	2.96×10^{-15}	8.90×10^{-25}
e	4.82×10^{-6}	3.75×10^{-9}	3.16×10^{-24}	1.90×10^{-36}
f	0.085	0.378	8.48×10^{-18}	2.05×10^{-25}
g	2.53×10^{-4}	2.29×10^{-6}	2.51×10^{-13}	2.07×10^{-24}
h	9.56×10^{-5}	0.095	6.65×10^{-6}	4.21×10^{-4}
i	0.123	0.136	9.62×10^{-5}	5.25×10^{-6}

Conclusions

The large number of replicas confirms that the predicted absolute binding free energies using the ESMACS protocol are definitively non-normal for the ligand-protein complexes investigated within the current study.

Acknowledgments

The authors are grateful to Dr. Alexander D. Wade for providing analysis script and for useful discussions.

References

(1) 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* (6), 20200007. DOI: 10.1098/rsfs.2020.0007.

(2) Wan, S.; Sinclair, R. C.; Coveney, P. V. Uncertainty quantification in classical molecular dynamics. *Philos Trans A Math Phys Eng Sci* **2021**, *379* (2197), 20200082. DOI: 10.1098/rsta.2020.0082 (acccessed 2021/06/11).

(3) Paketuryte, V.; Petrauskas, V.; Zubriene, A.; Abian, O.; Bastos, M.; Chen, W. Y.; Moreno, M. J.; Krainer, G.; Linkuviene, V.; Sedivy, A.; et al. Uncertainty in protein-ligand binding constants: asymmetric confidence intervals versus standard errors. *Eur Biophys J* **2021**, *50* (3-4), 661-670. DOI: 10.1007/s00249-021-01518-4

(4) Wan, S.; Bhati, A. P.; Wright, D. W.; Wall, I. D.; Graves, A. P.; Green, D.; Coveney, P. V. Ensemble Simulations and Experimental Free Energy Distributions: Evaluation and Characterization of Isoxazole Amides as SMYD3 Inhibitors. *J Chem Inf Model* **2022**, *62* (10), 2561-2570. DOI: 10.1021/acs.jcim.2c00255.

(5) Knapp, B.; Ospina, L.; Deane, C. M. Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas. *J Chem Theory Comput* **2018**, *14* (12), 6127-6138. DOI: 10.1021/acs.jctc.8b00391.

(6) 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* (8), 3696-3713. DOI: 10.1021/acs.jctc.5b00255.

(7) Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I. Y.; Berryman, J. T.; Brozell, S. R.; Cerutti, D. S.; T.E. Cheatham, I.; Cisneros, G. A.; Cruzeiro, V. W. D.; et al. *Amber 2022*; 2022.

(8) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Henin, J.; Jiang, W.; et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **2020**, *153* (4), 044130. DOI: 10.1063/5.0014475.