DeltaGzip: Computing Biopolymer-Ligand Binding Affinity via Kolmogorov Complexity and Lossless Compression

Tao Liu and Lena Simine*

Department of Chemistry, McGill University, Montreal, Quebec, H3A 0B8, Canada

The design of bio-sequences for biosensing and therapeutics is a challenging multi-step search and optimization task. In principle, computational modeling may speed up the design process by virtual screening of sequences based on their binding affinities to target molecules. However, in practice, existing machine-learned models trained to predict binding affinities lack the flexibility with respect to reaction conditions, and molecular dynamics simulations that can incorporate reaction conditions suffer from high computational costs. Here, we describe a computational approach called DeltaGzip that evaluates the free energy of binding in biopolymer-ligand complexes from ultra-short equilibrium molecular dynamics simulations. The entropy of binding is evaluated using the Kolmogorov complexity definition of entropy and approximated using a lossless compression algorithm, Gzip. We benchmark the method on a well-studied dataset of protein-ligand complexes comparing the predictions of DeltaGzip to the free energies of binding obtained using the Jarzynski equality and experimental measurements.

ORCID: orcid.org/0000-0002-1082-5570; orcid.org/0000-0002-8188-0550
Author email: tao.liu7@mail.mcgill.ca
Corresponding author email: lena.simine@mcgill.ca

Introduction

Short biopolymers such as peptide and nucleic acid aptamers have shown a capacity to bind diverse molecular targets strongly and selectively positioning them as promising candidates for therapeutics and biosensing. Examples of molecular targets include clinical drugs[1–5], food toxins[6–8], hormones[9–11], cells[12–14], etc. As illustrated in Figure 1A, a typical sequence design starts with a large library of fixed-length random (peptide or nucleic acids) sequences, followed by an iterative selection process in lab, such as SELEX (systematic evolution of ligands by exponential enrichment) for aptamers.[15–18] The resulting candidate sequences often require further optimization to improve their binding affinity, selectivity, and responsiveness.

The first task on the optimization list is improving the binding to target. It requires experimental evaluation of the binding affinity for many sequences. This is expensive and performing some of the search computationally would be an attractive option if a fast, accurate, and flexible enough computational method was available. A machine learning model trained on experimental data would have been the best solution, but data-driven approaches[19–29] are largely ineffective in this space. This is because the reaction conditions as well as measurement techniques vary depending on the intended applications and the measured values of binding affinity vary sensitively with them. Well-curated sufficiently large and diverse datasets with conditions and measurement techniques either standardized or accurately labeled for biopolymer-ligand complexes currently are unavailable to the research community.

In the absence of data, molecular dynamics (MD) simulations may be used to include custom reaction conditions in the calculation, including pH, temperature, ionic strength, choices of ions, etc. Several popular protocols for evaluating binding free energy using MD simulations have been developed, such as alchemical method[30], Jarzynski equality[31], and umbrella sampling method[32], which leverage a large number of trial paths, either virtual or physical paths, linking bound state and free state and serving as a hypothesized reaction coordinate, as shown in Figure 1B. Recent advances demonstrate improved accuracy of these techniques thanks to the development of machine learned potentials[33] and active learning[34]. These simulation protocols are computationally intensive, require hyper-parameterization, and do not present an adequate solution to the problem of computational high-throughput search in the sequence space of biopolymers for strong binders to a particular ligand.

In this paper, we present a cost-effective method for fast computational evaluation of the binding free energy in biopolymer-ligand complexes from MD simulations. Our method is based on the recently proposed approach to approximating the entropy in molecular systems by approximating the Kolmogorov complexity using lossless compression[35–40]. This approach removes the need for simulating the state-state transformation and thereby significantly reduces the required computational effort relative to existing methods. We call the method DeltaGzip to indicate that we evaluate the free energy using the popular lossless compression algorithm Gzip.[41] We benchmarked it against results obtained using the Jarzynski equality as well as experimental data and obtained strong results at a very low computational cost.

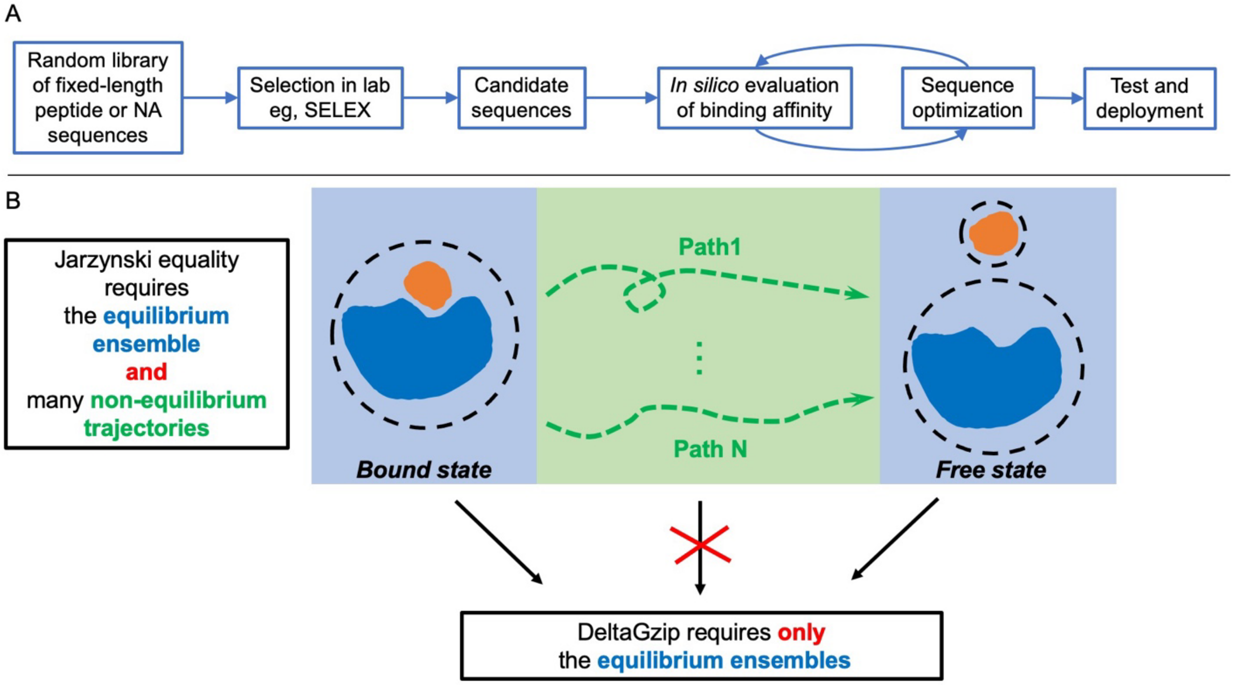
A
Random library of fixed-length peptide or NA sequences
Selection in lab eg, SELEX
Candidate sequences
In silico evaluation of binding affinity
Sequence optimization
Test and deployment
B
Jarzynski equality requires the equilibrium ensemble and many non-equilibrium trajectories
Path1
Path N
Bound state
Free state
DeltaGzip requires only the equilibrium ensembles

Figure 1. A) A typical computer-aided pipeline of designing biopolymer sequences for binding a molecular target of interest in applications such as biosensing. B) A schematic that highlights the computational advantage of DeltaGzip over other computational protocols for evaluating binding free energy from molecular simulations that commonly rely on simulating a large number of paths, either physical or virtual, connecting the bound and the free states. In contrast, the DeltaGzip only requires short equilibrium simulations of the terminal states.

Methods

Section I: Theoretical background

We evaluate the free energy change (ΔG) by evaluating the enthalpy change (ΔH) and the entropy change (ΔS) separately and combining them using the standard thermodynamic equality

$$\varDelta G = \varDelta H - T\varDelta S, \tag{1}$$

where T is temperature. To approximate the enthalpy of binding ΔH for a biopolymer-ligand complex, we subtract the energy stored in the nonbonded interactions (Lennard-Jones and Coulomb based on the force-field parameters, see Section III for details) between the biopolymer and ligand in the bound state from the energy stored in the unbound state (the latter vanishes by construction) and average over all configurations in our ensemble. The screening effect of the solvent is included via the medium dielectric constant in the Coulomb interaction term, and the enthalpy change within biopolymer is neglected as simplifying approximations.

Entropy is a fundamental concept in both thermodynamics and information theory. In thermodynamics, Gibbs entropy S_{Gibbs} of a macrostate χ described by macroscopic parameters such as temperature, volume, or pressure is given by Equation 2 where $P(x)$ is the probability of each microstate configuration x

$$S_{\text{Gibbs}} = -k_B \sum\nolimits_{x \in \chi} P(x) \ln P(x) \tag{2}$$

Similarly, in information theory, the Shannon's entropy[42] is defined for a random variable x with known sample space χ and probability distribution $P(x)$ as

$$H(P) = -\sum\nolimits_{x \in \chi} P(x) \log_2 P(x). \tag{3}$$

The relationship between the two quantities is given by

$$S_{\text{Gibbs}} = k_B \ln(2)\, H(P). \tag{4}$$

Unlike $H(P)$ that is defined for a random variable, the Kolmogorov complexity $K(x)$ is defined for the realizations of a random variable as the length of the shortest computer program that generates the data[43]. It has been shown that an inequality governs the relationship between Shannon's entropy and the Kolmogorov complexity[44]

$$H(P) \leq \sum_{x \in \chi} P(x) K(x)\,. \tag{5}$$

Combining $Eqn. 4$ and $Eqn. 5$, we see that using Kolmogorov complexity $K(x)$ we approach the Gibbs entropy S_{Gibbs} from above

$$S_{\mathrm{Gibbs}} \leq k_B \ln(2) \left(\sum_{x \in \chi} P(x) K(x) \right). \tag{6}$$

This inequality informs the strategy of choosing the compression algorithm that gives the lowest value of entropy on a given dataset. Gzip was used in this work. Overall, in order to approximate S_{Gibbs} we will draw samples of molecular configurations from the equilibrium Boltzmann distribution using a molecular dynamics simulation, approximate the Kolmogorov complexity for each sample using a lossless compression algorithm Gzip, and take an average, see Section III for more details.

Section II: Choice of test systems

Our main goal in this paper is to demonstrate the performance of DeltaGzip. An existing limitation of this method is that the outcomes for different systems should be compared to each other only if the biopolymers have similar lengths. This does not pose any problems for the field of aptamer design (our target application) because the length of the biopolymer is typically fixed in the relevant high-throughput experiments. Nonetheless, it posed an additional constraint on the datasets we could use for benchmarking. We searched for a dataset in which the free energy was characterized under identical reaction conditions for all points in the dataset and in which the biopolymer length did not vary significantly. We have chosen to test the method on the dataset of 23 ligand-AmpC β-lactamase protein pairs where the binding free energy was determined from the inhibition constant K_i obtained from IC_{50} plots assuming competitive inhibition[45–50]. Importantly, the experimental protocol did not change from ligand to ligand. The binding affinity in this set of systems was explored computationally in the past using the Jarzynski equality with high reported correlation between the computational and experimental results. The results and the full details on the protein and the ligands can be found in the original publication[51].

Section III: Implementation

Although theoretically incomputable, Kolmogorov complexity can be approximated by the length of byte-stream obtained by lossless compression of the input. In this work, we chose Gzip as the lossless compressor for its simplicity and easy implementation, implemented as part of Python's native module. It has been used for this purpose in the past[52–54] and preferred over

https://doi.org/10.26434/chemrxiv-2024-cq702 **ORCID:** https://orcid.org/0000-0002-8188-0550 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC-ND 4.0

other common algorithms such as LZMA (Lempel–Ziv–Markov chain algorithm) and Bzip2 for its speed advantage.

Representation: When compressing data that represents a molecular configuration a choice of representation must be made. The entropy change that will be captured by compression depends sensitively on the representation. For example, the Gzip compressor would struggle to distinguish two distinct states by $\Delta S_{\mathrm{Gibbs}}$ if provided with too general or too detailed information about the configurations in each state, or alternatively it could yield a non-vanishing $\Delta S_{\mathrm{Gibbs}}$ for the same state even though the system has simply shifted and rotated in space. A good representation would be invariant to transformations that do not alter the physics (translations and rotations of the entire system), and it would include important details while omitting the rest thereby striking a good balance between including useful information and making the compression task easier.

In our work, we deal with two macrostates of a biopolymer: bound state and free state. Since the molecular ligands that we explore are small molecules that are not expected to significantly contribute to the entropy change we simplify the calculation by only including the biopolymer backbone degrees of freedom (DOFs). To ensure that our representation is roto-translationally invariant we convert the Cartesian coordinates of atoms that belong to the biopolymer backbone into the Z-matrix representation of internal coordinates. Each internal coordinate is rounded to the first decimal point (bond length is measured in angstrom, and angle and dihedral are in radian), multiplied by 10 and stored as an integer.

As a preliminary analysis we scatter the entropy in the bound state against the free state evaluated for each DOF separately, see Figure 2A. To do this, we collect the values for each DOF separately from an equilibrium MD simulation of bound/free state into 1D arrays and compress them individually using Gzip. By scattering the compressed sizes of these arrays for the bound and the free states against each other, we notice that for a subset of DOFs shown in orange in Figure 2A the change in entropy is very small ($\leq$ 125 bytes) in going from the bound to the free state. We label these DOFs 'unresponsive DOF'. DOFs which show a stronger entropy change ($>$ 125 bytes) shown in green in Figure 2A are labelled 'responsive DOF'. Figure 2B shows the backbone atoms that exclusively belong to the 'responsive' (green) and 'unresponsive' (orange) DOFs on the protein structure. The atoms in gray participate in both responsive and unresponsive DOFs. In an effort to make the compression task easier we will exclude these 'unresponsive DOFs' from the calculation of entropy for each full molecular configuration.

Entropy evaluation: To summarize, the coordinates of backbone atoms in each frame generated by the equilibrium MD simulations are transformed into internal coordinates, 'unresponsive' DOFs are removed, each value is rounded to the first decimal point (bond length is measured in angstrom, and angle and dihedral are in radian), multiplied by 10 and stored as an integer, the result is compressed using the Gzip compression algorithm. The sizes of compressed files are summed together, divided by the number of samples in the ensemble, and the result is plugged into $Eqn. 6$ producing the value of entropy in physical units.

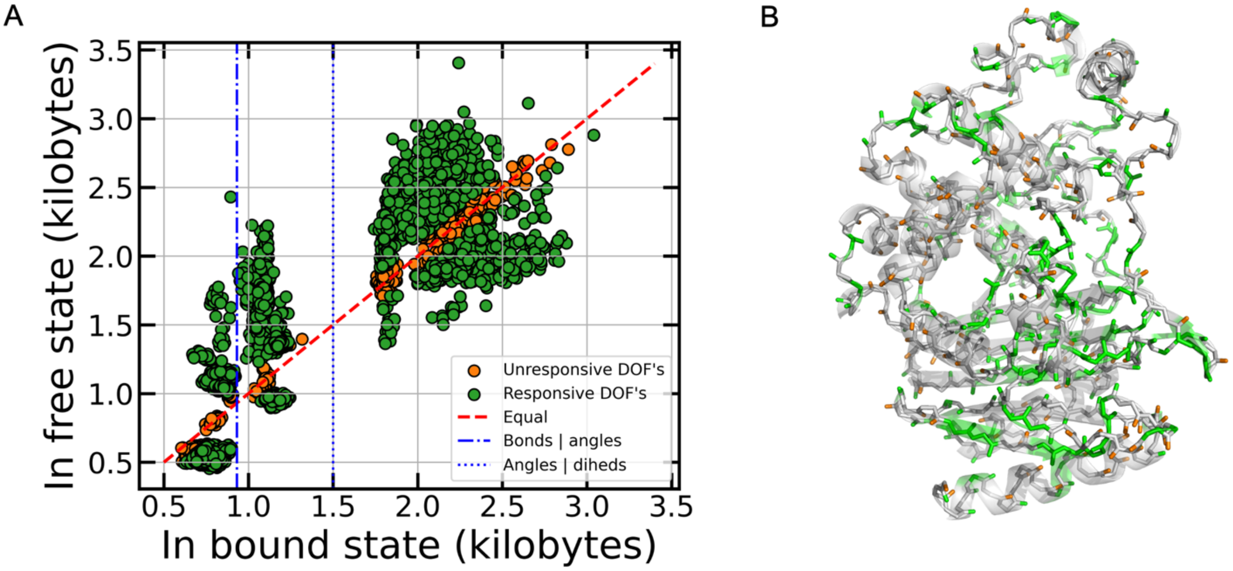

Figure 2. Identifying the subset of degrees of freedom (DOFs) to be used in compression. Panel A) Scatter plot of compressed DOFs in the free state against the compressed DOFs in the bound state. DOFs that are close to the diagonal (orange) are compressed to very similar sizes in both states ('unresponsive DOFs'), whereas DOFs away from the diagonal (green) show a difference in compressed size ('responsive DOFs'). Vertical lines indicate boundaries between regions characterized by a different type of DOFs: bonds and angles, and angles and dihedrals. Panel B) highlights the backbone atoms exclusively involved in responsive DOFs (green) vs. in unresponsive DOFs (orange). The gray regions indicate the backbone atoms that participate in both responsive and unresponsive DOFs.

MD simulations: In this work, we run both equilibrium and non-equilibrium molecular dynamics (MD) simulations using OpenMM[55]. Amber14 force fields are used for modeling biopolymers (FF14SB for protein)[56,57], general Amber Force Field (GAFF) 2.11 for molecular target[58], and TIP3P model[59] for water and ions, unless indicated otherwise. Langevin integrator is applied with friction coefficient of 1 per ps and time-step of 1fs. Particle Mesh Ewald (PME) method is used with 1nm cutoff and fractional error tolerance of 5.0×10^{-4} in force computation for modeling electrostatic interactions. Proteins are simulated in 300K, neutralized by adding 3 Cl^- ions to the water box and accompanied by no other ions, consistent with the computational work reported in literature.[51] The simulations in the bound state are initialized with the biopolymer-ligand complex and to initialize the free state simulation the ligand is removed.

In our preliminary calculations the equilibrium simulations consisted of 1ns equilibration, 20ns sampling, generating 5000 configurations (4ps print interval). By systematically cutting short the trajectory used in our modeling, we found that ultra-short trajectories of only 100ps of sampling for the bound and free states were sufficient giving similar accuracy to longer trajectories and making this protocol extremely computationally efficient. The results shown in Figure 3 were obtained using 100ps-long trajectories.

Jarzynski equality calculations: In non-equilibrium MD simulations, pulling is realized by displacing an external harmonic force, force constant $k = 20000\ kJ/mol/nm^2$ or $16.6\ pN/\text{Å}$, which is exerted on the molecular target in the bound state structure. Molecular target is pulled along a direction that was chosen in a way that leads to the least clash with biopolymer. A total pulling distance of 1nm is evenly divided into 3200 stops and the external force recurrently snaps to the next stop after being fixed for 50 steps (1fs time step). So total simulation time is 160ps. In this work, for each biopolymer-ligand pair, a total of 5000 pulling trajectories were simulated with initial configurations sampled from an NVT ensemble simulation in which the external force is fixed at the location of the bound molecular target, so that the Hamiltonian in the initial configuration-generating simulation is the same as that at the start of each pulling simulation, as required by Jarzynski equality theory.

Results and discussion

The DeltaGzip protocol presented in the methods section aims to offer a cost-effective computational evaluation of binding free energy in biopolymer–ligand systems in a way that can incorporate realistic bioassay conditions in a flexible manner. We demonstrate a protocol for free energy calculations from very short molecular dynamics simulations based on the Kolmogorov complexity approximated using lossless compression. This approach circumvents the need for lengthy sampling simulations and offers a much cheaper and faster simulation option while maintaining the flexibility in specifying reaction conditions and producing high-accuracy results. We note that in the context of computational screening our approach should be used for the comparative analysis of biopolymers of similar lengths. This scenario arises in many important application cases such as the design of aptamers for biosensing and therapeutics. We leave the generalization to length-independent evaluation of free energy to future work.

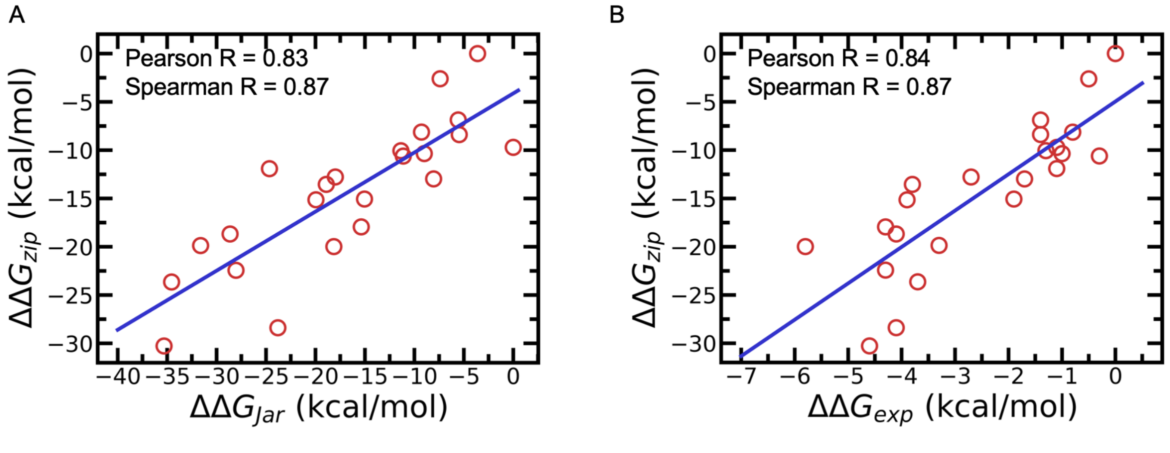

Figure 3. Correlation between ΔG_{zip} and reference free energy: A) ΔG_{zip} vs ΔG_{jar} – the free energy evaluated computationally using the Jarzynski equality, B) ΔG_{zip}vs ΔG_{exp} – experimentally evaluated free energy of binding. For simplicity, the plot shows the relative free energy change $\Delta\Delta G = \Delta G - \Delta G_{max}$.

Figure 3 shows the performance of DeltaGzip on 23 ligand-AmpC β-lactamase protein pairs. The performance against computational results obtained using the Jarzynski equality is shown in Figure 3A: the correlation between the two computational techniques is 83% linear and 87% Spearman correlations. We found that the results of the Jarzynski equality calculation depended sensitively on the choices made in the implementation such as the direction of pulling and the type of equilibration we performed. We attribute some of the lost correlation ultimately to the compromises we had to make in the Jarzynski equality implementation in order to keep the calculation affordable. The performance on the experimental dataset shown in Figure 3B is very encouraging. DeltaGzip achieves 84% linear correlation with the experimentally measured free energy of binding, and 87% Spearman correlation that indicates the correct ranking of binding strengths. This result suggests that the method can be useful and practical in computational screening of biopolymer-ligand complexes in scenarios in which data is limited and/or the flexibility with respect to reaction conditions is of essence.

Conclusions

In this work, we proposed and tested a protocol called DeltaGzip for evaluating biopolymer-target binding free energy from MD simulation using Kolmogorov complexity. We showed that this approach has achieved strong correlations with reference data (Figure 3) on a dataset of protein-ligand complexes and we recommend its application for computational screening of biopolymer-ligand pairs (biopolymers must be of similar lengths) for important applications such as therapeutics and biosensing.

Author information and author contributions:
TL executed the research, wrote the manuscript; LS planned the research, contributed to the writing of the manuscript.

Note:
The authors declare no competing financial interest.

Acknowledgement:
We acknowledge support from the AI for Design Challenge Program from National Research Council (NRC), IVADO, and Natural Sciences and Engineering Research Council of Canada (NSERC) via the NOVA Grant program. T.L. also acknowledges the doctoral fellowship from Fonds de recherche du Québec – Nature et technologies (FRQNT).

References

(1) Neves, M. A. D.; Reinstein, O.; Saad, M.; Johnson, P. E. Defining the Secondary Structural Requirements of a Cocaine-Binding Aptamer by a Thermodynamic and Mutation Study. *Biophysical Chemistry* **2010**, *153* (1), 9–16. https://doi.org/10.1016/j.bpc.2010.09.009.

(2) Neves, M. A. D.; Reinstein, O.; Johnson, P. E. Defining a Stem Length-Dependent Binding Mechanism for the Cocaine-Binding Aptamer. A Combined NMR and Calorimetry Study. *Biochemistry* **2010**, *49* (39), 8478–8487. https://doi.org/10.1021/bi100952k.

(3) Reinstein, O.; Yoo, M.; Han, C.; Palmo, T.; Beckham, S. A.; Wilce, M. C. J.; Johnson, P. E. Quinine Binding by the Cocaine-Binding Aptamer. Thermodynamic and Hydrodynamic Analysis of High-Affinity Binding of an Off-Target Ligand. *Biochemistry* **2013**, *52* (48), 8652–8662. https://doi.org/10.1021/bi4010039.

(4) Li, H.; Dauphin-Ducharme, P.; Ortega, G.; Plaxco, K. W. Calibration-Free Electrochemical Biosensors Supporting Accurate Molecular Measurements Directly in Undiluted Whole Blood. *J. Am. Chem. Soc.* **2017**, *139* (32), 11207–11213. https://doi.org/10.1021/jacs.7b05412.

(5) Dauphin-Ducharme, P.; Yang, K.; Arroyo-Currás, N.; Ploense, K. L.; Zhang, Y.; Gerson, J.; Kurnik, M.; Kippin, T. E.; Stojanovic, M. N.; Plaxco, K. W. Electrochemical Aptamer-Based Sensors for Improved Therapeutic Drug Monitoring and High-Precision, Feedback-Controlled Drug Delivery. *ACS Sens.* **2019**, *4* (10), 2832–2837. https://doi.org/10.1021/acssensors.9b01616.

(6) McKeague, M.; Velu, R.; Hill, K.; Bardóczy, V.; Mészáros, T.; DeRosa, M. Selection and Characterization of a Novel DNA Aptamer for Label-Free Fluorescence Biosensing of Ochratoxin A. *Toxins* **2014**, *6* (8), 2435–2452. https://doi.org/10.3390/toxins6082435.

(7) De Girolamo, A.; McKeague, M.; Miller, J. D.; DeRosa, M. C.; Visconti, A. Determination of Ochratoxin A in Wheat after Clean-up through a DNA Aptamer-Based Solid Phase Extraction Column. *Food Chemistry* **2011**, *127* (3), 1378–1384. https://doi.org/10.1016/j.foodchem.2011.01.107.

(8) Costantini, F.; Sberna, C.; Petrucci, G.; Reverberi, M.; Domenici, F.; Fanelli, C.; Manetti, C.; De Cesare, G.; DeRosa, M.; Nascetti, A.; Caputo, D. Aptamer-Based Sandwich Assay for on Chip Detection of Ochratoxin A by an Array of Amorphous Silicon Photosensors. *Sensors and Actuators B: Chemical* **2016**, *230*, 31–39. https://doi.org/10.1016/j.snb.2016.02.036.

(9) Contreras Jiménez, G.; Eissa, S.; Ng, A.; Alhadrami, H.; Zourob, M.; Siaj, M. Aptamer-Based Label-Free Impedimetric Biosensor for Detection of Progesterone. *Anal. Chem.* **2015**, *87* (2), 1075–1082. https://doi.org/10.1021/ac503639s.

(10) Ye, C.; Wang, M.; Min, J.; Tay, R. Y.; Lukas, H.; Sempionatto, J. R.; Li, J.; Xu, C.; Gao, W. A Wearable Aptamer Nanobiosensor for Non-Invasive Female Hormone Monitoring. *Nat. Nanotechnol.* **2023**. https://doi.org/10.1038/s41565-023-01513-0.

(11) Feng, S.; Chen, C.; Song, C.; Ding, X.; Wang, W.; Que, L. Optical Aptamer-Based Sensors for Detecting Plant Hormones. *IEEE Sensors J.* **2021**, *21* (5), 5743–5750. https://doi.org/10.1109/JSEN.2020.3041266.

(12) Xu, Y.; Phillips, J. A.; Yan, J.; Li, Q.; Fan, Z. H.; Tan, W. Aptamer-Based Microfluidic Device for Enrichment, Sorting, and Detection of Multiple Cancer Cells. *Anal. Chem.* **2009**, *81* (17), 7436–7442. https://doi.org/10.1021/ac9012072.

(13) Medley, C. D.; Bamrungsap, S.; Tan, W.; Smith, J. E. Aptamer-Conjugated Nanoparticles for Cancer Cell Detection. *Anal. Chem.* **2011**, *83* (3), 727–734. https://doi.org/10.1021/ac102263v.

(14) Herr, J. K.; Smith, J. E.; Medley, C. D.; Shangguan, D.; Tan, W. Aptamer-Conjugated Nanoparticles for Selective Collection and Detection of Cancer Cells. *Anal. Chem.* **2006**, *78* (9), 2918–2924. https://doi.org/10.1021/ac052015r.

(15) Yang, K.-A.; Pei, R.; Stojanovic, M. N. In Vitro Selection and Amplification Protocols for Isolation of Aptameric Sensors for Small Molecules. *Methods* **2016**, *106*, 58–65. https://doi.org/10.1016/j.ymeth.2016.04.032.

(16) Huang, P.-J. J.; Liu, J. A DNA Aptamer for Theophylline with Ultrahigh Selectivity Reminiscent of the Classic RNA Aptamer. *ACS Chem. Biol.* **2022**, *17* (8), 2121–2129. https://doi.org/10.1021/acschembio.2c00179.

(17) Huang, P.-J. J.; Liu, J. Selection of Aptamers for Sensing Caffeine and Discrimination of Its Three Single Demethylated Analogues. *Anal. Chem.* **2022**, *94* (7), 3142–3149. https://doi.org/10.1021/acs.analchem.1c04349.

(18) Shoara, A. A.; Slavkovic, S.; Donaldson, L. W.; Johnson, P. E. Analysis of the Interaction between the Cocaine-Binding Aptamer and Its Ligands Using Fluorescence Spectroscopy. *Can. J. Chem.* **2017**, *95* (12), 1253–1260. https://doi.org/10.1139/cjc-2017-0380.

(19) Meng, Z.; Xia, K. Persistent Spectral–Based Machine Learning (PerSpect ML) for Protein-Ligand Binding Affinity Prediction. *Sci. Adv.* **2021**, *7* (19), eabc5329. https://doi.org/10.1126/sciadv.abc5329.

(20) Rube, H. T.; Rastogi, C.; Feng, S.; Kribelbauer, J. F.; Li, A.; Becerra, B.; Melo, L. A. N.; Do, B. V.; Li, X.; Adam, H. H.; Shah, N. H.; Mann, R. S.; Bussemaker, H. J. Prediction of Protein–Ligand Binding Affinity from Sequencing Data with Interpretable Machine Learning. *Nat Biotechnol* **2022**, *40* (10), 1520–1527. https://doi.org/10.1038/s41587-022-01307-0.

(21) Wang, K.; Zhou, R.; Li, Y.; Li, M. DeepDTAF: A Deep Learning Method to Predict Protein–Ligand Binding Affinity. *Briefings in Bioinformatics* **2021**, *22* (5), bbab072. https://doi.org/10.1093/bib/bbab072.

(22) Ahmed, A.; Mam, B.; Sowdhamini, R. DEELIG: A Deep Learning Approach to Predict Protein-Ligand Binding Affinity. *Bioinform Biol Insights* **2021**, *15*, 117793222110303. https://doi.org/10.1177/11779322211030364.

(23) Lu, W.; Wu, Q.; Zhang, J.; Rao, J.; Li, C.; Zheng, S. *TANKBind: Trigonometry-Aware Neural NetworKs for Drug-Protein Binding Structure Prediction*; preprint; Biophysics, 2022. https://doi.org/10.1101/2022.06.06.495043.

(24) Somnath, V. R.; Bunne, C.; Krause, A. Multi-Scale Representation Learning on Proteins. arXiv April 4, 2022. http://arxiv.org/abs/2204.02337 (accessed 2024-03-04).

(25) Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: Deep Drug–Target Binding Affinity Prediction. *Bioinformatics* **2018**, *34* (17), i821–i829. https://doi.org/10.1093/bioinformatics/bty593.

(26) Shen, C.; Zhang, X.; Deng, Y.; Gao, J.; Wang, D.; Xu, L.; Pan, P.; Hou, T.; Kang, Y. Boosting Protein–Ligand Binding Pose Prediction and Virtual Screening Based on Residue–Atom Distance Likelihood Potential and Graph Transformer. *J. Med. Chem.* **2022**, *65* (15), 10691–10706. https://doi.org/10.1021/acs.jmedchem.2c00991.

(27) Moon, S.; Zhung, W.; Yang, S.; Lim, J.; Kim, W. Y. PIGNet: A Physics-Informed Deep Learning Model toward Generalized Drug–Target Interaction Predictions. *Chem. Sci.* **2022**, *13* (13), 3661–3673. https://doi.org/10.1039/D1SC06946B.

(28) Jiang, D.; Hsieh, C.-Y.; Wu, Z.; Kang, Y.; Wang, J.; Wang, E.; Liao, B.; Shen, C.; Xu, L.; Wu, J.; Cao, D.; Hou, T. InteractionGraphNet: A Novel and Efficient Deep Graph Representation Learning Framework for Accurate Protein–Ligand Interaction Predictions. *J. Med. Chem.* **2021**, *64* (24), 18209–18232. https://doi.org/10.1021/acs.jmedchem.1c01830.

(29) Ballester, P. J.; Mitchell, J. B. O. A Machine Learning Approach to Predicting Protein–Ligand Binding Affinity with Applications to Molecular Docking. *Bioinformatics* **2010**, *26* (9), 1169–1175. https://doi.org/10.1093/bioinformatics/btq112.

(30) Tembre, B. L.; Mc Cammon, J. A. Ligand-Receptor Interactions. *Computers & Chemistry* **1984**, *8* (4), 281–283. https://doi.org/10.1016/0097-8485(84)85020-2.

(31) Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **1997**, *78* (14), 2690–2693. https://doi.org/10.1103/PhysRevLett.78.2690.

(32) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *Journal of Computational Physics* **1977**, *23* (2), 187–199. https://doi.org/10.1016/0021-9991(77)90121-8.

(33) Sabanés Zariquiey, F.; Galvelis, R.; Gallicchio, E.; Chodera, J. D.; Markland, T. E.; De Fabritiis, G. Enhancing Protein–Ligand Binding Affinity Predictions Using Neural Network Potentials. *J. Chem. Inf. Model.* **2024**, acs.jcim.3c02031. https://doi.org/10.1021/acs.jcim.3c02031.

(34) Koby, S. B.; Gutkin, E.; Gusev, F.; Kottke, C.; Patel, S.; Isayev, O.; Kurnikova, M. G. Optimizing High-Throughput Binding Free Energy Simulations for Small Molecule Drug Discovery. *Biophysical Journal* **2024**, *123* (3), 296a. https://doi.org/10.1016/j.bpj.2023.11.1846.

(35) Avinery, R.; Kornreich, M.; Beck, R. Universal and Accessible Entropy Estimation Using a Compression Algorithm. *Phys. Rev. Lett.* **2019**, *123* (17), 178102. https://doi.org/10.1103/PhysRevLett.123.178102.

(36) Martiniani, S.; Lemberg, Y.; Chaikin, P. M.; Levine, D. Correlation Lengths in the Language of Computable Information. *Phys. Rev. Lett.* **2020**, *125* (17), 170601. https://doi.org/10.1103/PhysRevLett.125.170601.

(37) Martiniani, S.; Chaikin, P. M.; Levine, D. Quantifying Hidden Order out of Equilibrium. *Phys. Rev. X* **2019**, *9* (1), 011031. https://doi.org/10.1103/PhysRevX.9.011031.

(38) Melchert, O.; Hartmann, A. K. Analysis of the Phase Transition in the Two-Dimensional Ising Ferromagnet Using a Lempel-Ziv String-Parsing Scheme and Black-Box Data-Compression Utilities. *Phys. Rev. E* **2015**, *91* (2), 023306. https://doi.org/10.1103/PhysRevE.91.023306.

(39) Vogel, E. E.; Saravia, G.; Cortez, L. V. Data Compressor Designed to Improve Recognition of Magnetic Phases. *Physica A: Statistical Mechanics and its Applications* **2012**, *391* (4), 1591–1601. https://doi.org/10.1016/j.physa.2011.09.005.

(40) Vogel, E. E.; Saravia, G.; Ramirez-Pastor, A. J. Phase Transitions in a System of Long Rods on Two-Dimensional Lattices by Means of Information Theory. *Phys. Rev. E* **2017**, *96* (6), 062133. https://doi.org/10.1103/PhysRevE.96.062133.

(41) Gailly, J. GNU Gzip, 1992.

(42) Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* **1948**, *27* (3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

(43) Kolmogorov, A. N. Three Approaches to the Quantitative Definition of Information. *Problems of information transmission* **1965**, *1* (1), 1–7.

(44) Li, M.; Vitányi, P. *An Introduction to Kolmogorov Complexity and Its Applications*; Texts in Computer Science; Springer New York: New York, NY, 2008. https://doi.org/10.1007/978-0-387-49820-1.

(45) Tondi, D.; Morandi, F.; Bonnet, R.; Costi, M. P.; Shoichet, B. K. Structure-Based Optimization of a Non-β-Lactam Lead Results in Inhibitors That Do Not Up-Regulate β-Lactamase Expression in Cell Culture. *J. Am. Chem. Soc.* **2005**, *127* (13), 4632–4639. https://doi.org/10.1021/ja042984o.

(46) Babaoglu, K.; Simeonov, A.; Irwin, J. J.; Nelson, M. E.; Feng, B.; Thomas, C. J.; Cancian, L.; Costi, M. P.; Maltby, D. A.; Jadhav, A.; Inglese, J.; Austin, C. P.; Shoichet, B. K. Comprehensive Mechanistic Analysis of Hits from High-Throughput and Docking Screens against β-Lactamase. *J. Med. Chem.* **2008**, *51* (8), 2502–2511. https://doi.org/10.1021/jm701500e.

(47) Powers, R. A.; Morandi, F.; Shoichet, B. K. Structure-Based Discovery of a Novel, Noncovalent Inhibitor of AmpC β-Lactamase. *Structure* **2002**, *10* (7), 1013–1023. https://doi.org/10.1016/S0969-2126(02)00799-2.

(48) Barelier, S.; Eidam, O.; Fish, I.; Hollander, J.; Figaroa, F.; Nachane, R.; Irwin, J. J.; Shoichet, B. K.; Siegal, G. Increasing Chemical Space Coverage by Combining Empirical and Computational Fragment Screens. *ACS Chem. Biol.* **2014**, *9* (7), 1528–1535. https://doi.org/10.1021/cb5001636.

(49) Teotico, D. G.; Babaoglu, K.; Rocklin, G. J.; Ferreira, R. S.; Giannetti, A. M.; Shoichet, B. K. Docking for Fragment Inhibitors of AmpC β-Lactamase. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106* (18), 7455–7460. https://doi.org/10.1073/pnas.0813029106.

(50) Babaoglu, K.; Shoichet, B. K. Deconstructing Fragment-Based Inhibitor Discovery. *Nat Chem Biol* **2006**, *2* (12), 720–723. https://doi.org/10.1038/nchembio831.

(51) Truong, D. T.; Li, M. S. Probing the Binding Affinity by Jarzynski's Nonequilibrium Binding Free Energy and Rupture Time. *J. Phys. Chem. B* **2018**, *122* (17), 4693–4699. https://doi.org/10.1021/acs.jpcb.8b02137.

(52) Ehret, K. Kolmogorov Complexity of Morphs and Constructions in English. *LiLT* **2014**, *11*. https://doi.org/10.33011/lilt.v11i.1363.

(53) Ichimiya, M.; Nakamura, I. Randomness Representation in Turbulent Flows with Kolmogorov Complexity (In Mixing Layer). *JFST* **2013**, *8* (3), 407–422. https://doi.org/10.1299/jfst.8.407.

(54) Bloem, P.; Mota, F.; De Rooij, S.; Antunes, L.; Adriaans, P. A Safe Approximation for Kolmogorov Complexity. In *Algorithmic Learning Theory*; Auer, P., Clark, A., Zeugmann, T., Zilles, S., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2014; Vol. 8776, pp 336–350. https://doi.org/10.1007/978-3-319-11662-4_24.

(55) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L.-P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comput Biol* **2017**, *13* (7), e1005659. https://doi.org/10.1371/journal.pcbi.1005659.

(56) Galindo-Murillo, R.; Robertson, J. C.; Zgarbová, M.; Šponer, J.; Otyepka, M.; Jurečka, P.; Cheatham, T. E. Assessing the Current State of Amber Force Field Modifications for DNA. *J. Chem. Theory Comput.* **2016**, *12* (8), 4114–4127. https://doi.org/10.1021/acs.jctc.6b00186.

(57) 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. https://doi.org/10.1021/acs.jctc.5b00255.

(58) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J Comput Chem* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.

(59) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *The Journal of Chemical Physics* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.