DynamicBind: Predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model

Anonymous Author(s)
Affiliation
Address
email

Abstract

While significant advances have been made in predicting static protein structures, the inherent dynamics of proteins, modulated by ligands, are crucial for understanding protein function and facilitating drug discovery. Traditional docking methods, frequently used in studying protein-ligand interactions, typically treat proteins as rigid. While molecular dynamics simulations can propose appropriate protein conformations, they're computationally demanding due to rare transitions between biologically relevant equilibrium states. In this study, we present DynamicBind, a novel method that employs equivariant geometric diffusion networks to construct a smooth energy landscape, promoting efficient transitions between different equilibrium states. DynamicBind accurately recovers ligand-specific conformations from unbound protein structures without the need for holo-structures or extensive sampling. Our experiments reveal that DynamicBind can accommodate a wide range of large protein conformational changes and identify novel cryptic pockets in unseen protein targets. As a result, DynamicBind shows potential in accelerating the development of small molecules for previously undruggable targets and expanding the horizons of computational drug discovery.

1 Introduction

Remarkable progress has been achieved in the realm of protein structure prediction from sequence data, with AlphaFold leading the way in the prediction of nearly all structures in the human proteome [1–4]. However, these models generate a single static conformation for each protein sequence, despite the fact that proteins are inherently dynamic and generally adopt multiple conformations to perform their functions [5, 6]. The ability of proteins to interconvert between different conformations is central to their biological activities in all domains of life. The therapeutic effect of drug molecules arises from their specific binding to only some conformations of the target proteins and thereby modulating essential biological activities by altering the conformational landscape of these proteins [7–10]. In practice, nowadays the interactions between proteins and ligands are studied through molecular docking methods computationally. Docking is a key component of structure-based drug discovery [11]. Nevertheless, despite the widespread recognition of the importance of protein dynamics, traditional docking methods often treat proteins as rigid, or in some cases, as being only partially flexible, permitting only selected side-chains to move, to manage computational costs [12, 13]. This simplification leads to inferior performance in realistic scenarios where the input protein structures are in conformations distinct from the typically unavailable ligand-bounded holo-state conformations [14, 15].

Here, we present DynamicBind, an E(3)-equivariant diffusion-based deep generative model designed for 'dynamic docking'. Different from traditional docking methods, DynamicBind can efficiently

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

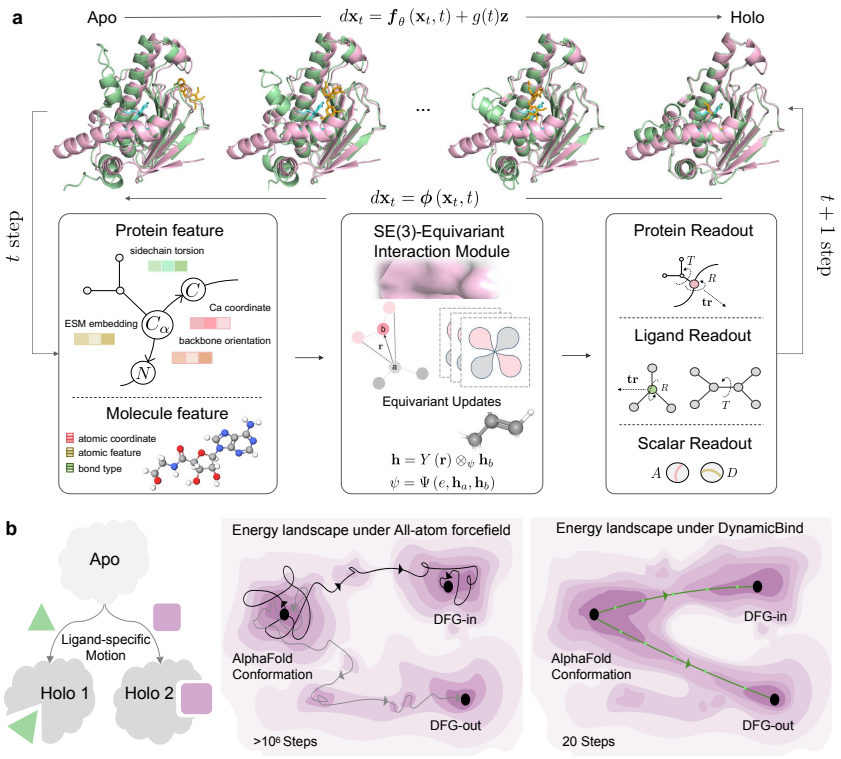

Figure 1: Overview of DynamicBind Model. **(a)** Diffusion and reverse generative processes are conducted between apo (without ligand bounded) and holo (with ligand bounded) structures. The holo state is represented in pink, the initial apo and the model-predicted conformation in green. The output readouts include the predicted updates: global translation and rotation for both the ligand and each protein residue, the rotation of torsional angles for the ligands and chi angles for the protein residues. Binding affinity and confidence score are also predicted. **(b)** When the protein binds with two different ligands, DynamicBind could predict the two different holo conformations within 20 generative steps, while millions of steps are needed for all-atom Molecular Dynamics simulations.

adjust the protein conformation from its initial AlphaFold prediction to a holo-like state. It is capable of handling a wide range of large conformational changes during prediction, such as the well-known DFG-in to DFG-out transition in kinase proteins, which is a formidable challenge for other methods [16, 17].

2 Method

DynamicBind is a diffusion-based generative model equipped with E(3)-equivariant interaction modules and coarse-grained protein features. As shown in Fig. 1(a), DynamicBind model learns to execute 'dynamic docking', a process that performs prediction of protein-ligand complex structures while accommodating substantial protein conformational changes.

During inference, DynamicBind receives apo-like protein structures (in the present study, conformations predicted by AlphaFold) and small molecule ligands as inputs, and randomly places the ligand around the protein initially. At each step, the features and the coordinates of the protein and the ligand are embedded by an E(3)-equivariant interaction module Over the course of $T = 20$ steps of reverse denoising generative process, the model gradually translates and rotates the ligand by adjusting its internal torsional angles, and simultaneously translates and rotates the protein residues and modifying the side-chain chi angles [18].

Canonical diffusion-based models is trained by perturbing the ground-truth data distribution with Gaussian noise of varying magnitudes in diffusion process and denoising in reverse generative

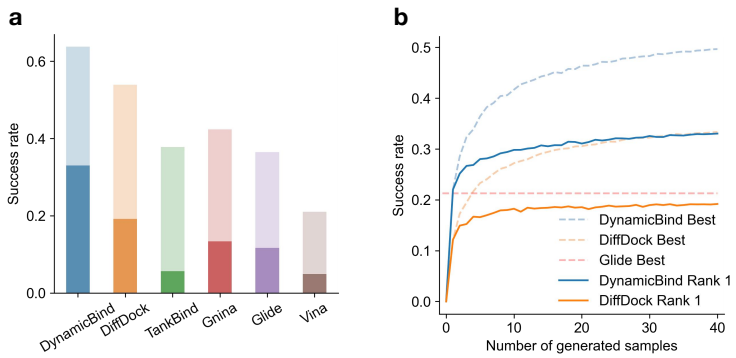

Figure 2: DynamicBind outperforms other methods in predicting ligand poses. **(a)** Dark and light shades represent success rates under stringent (ligand RMSD < 2Å, clash score < 0.35) and relaxed (ligand RMSD < 5Å, clash score < 0.5) criteria, respectively. **(b)** With the cLDDT score exploited as ranking measurement, the success rate of DynamicBind is enhanced.

process [19–22]. However, structures of proteins are highly constrained in many ways, e.g. residues are linked by peptide bonds and bond lengths are governed by chemical principles. Studies have demonstrated that, when decoys (i.e. structure P^t at step t) are generated using Gaussian noise, the model primarily learns only to revert to a chemically stable conformation [23], i.e. apo conformations predicted by AlphaFold in our task. It is challenge for the model to accurately predict long timescale transformations of biological relevance, which are our primary concern.

To cope with these challenges, our method employs a morph-like transformation to produce protein decoys, where the native conformation is gradually transitioned towards the AlphaFold-predicted conformation (see Sec. A in supplementary materials for detail). The decoys generated by our morph-like transformation generally satisfy the basic chemical constraints, allowing our model to concentrate on learning biophysically relevant state-changing events. Compared with the slow transitions between meta-stable states by unbiased molecular dynamics simulations, our method features a significantly more funneled energy landscape, effectively lowering the free energy barrier between biologically meaningful states, Fig. 1(b).

3 Experiments and Results

3.1 DynamicBind achieves higher success rate in ligand pose prediction

Our method is evaluated on PDBbind dataset [24] and a curated Major Drug Target (MDT) dataset. The MDT dataset consists of 599 structures that were deposited in or after 2020, with both drug-like ligands and proteins from four major protein families: kinases, GPCRs, nuclear receptors and io channels, which represent the targets of about 70% of FDA-approved small-molecule drugs [25]. In line with previous works [26, 27, 15], we trained the model with a chronological, time-based split on the PDBbind dataset (More in. A.3 and A.4).

Traditionally, models are evaluated by using the holo protein structures as input for ligand pose prediction. However, holo conformations exhibit strong shape and charge complementarity to co-crystallized ligands, which may simplify ligand pose prediction [7], but hard to obtain in practice. In this experiment, a more challenging and realistic scenario is adopted. We assume that the holo protein conformation is not available and only use the protein conformations predicted by AlphaFold as input for ligand pose prediction. Experimental results on both PDBbind and MDT test set are combined in Fig. 2 due to the space limit. Individual results are provided in the Supplementary Materials.

As a generative model, DynamicBind could sample multiple protein-ligand conformations, and the contact-LDDT (cLDDT, Sec. A.2) scoring module is designed to rank those sampled structures. With cLDDT exploited as ranking measurement, the success rate of DynamicBind is enhanced from 0.33 to 0.5, considerably outperforms DiffDock and the best force-field-based method, GLIDE (Fig. 2(b)). Given that GLIDE may generate different amount of samples for each case, we draw its best performance as a reference line.

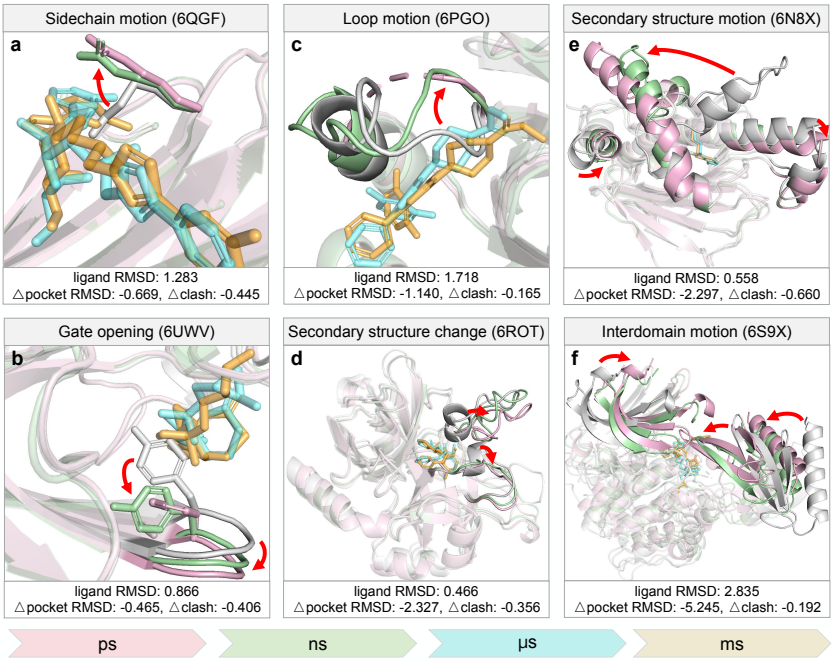

Figure 3: DynamicBind predicts protein conformational changes upon ligand binding, across a range of time scales, from picosecond to millisecond. Negative Δpocket RMSD and Δclash scores indicate that the predicted structures align better with the target crystal structures than the initial structures.

3.2 DynamicBind covers multi-scale protein conformation changes

To illustrate the capability of DynamicBind in predicting unique protein conformational changes upon ligand binding, Fig. 3 depicts six types of predicted protein conformational changes. All cases were identified from the PDBbind test set. The crystal structures, AlphaFold structures and our predicted structures are shown in pink, white and green. The native ligand poses and our predicted poses are shown in cyan and orange, respectively. In Fig. 3(a), a side-chain motion is executed by DynamicBind to avoid a clash which the initial AlphaFold structure may encounter. In Fig. 3(b), a gate opening is performed to make the pocket accessible, while the pocket is blocked by a Tyrosine in AlphaFold structure. In Fig. 3(c), a flexible loop is moved away to avoid of intersection with the ligand. In Fig. 3(d), alpha helices transform into loops near the ligand binding site. In Fig. 3(e), a substantial secondary structure motion is observed in the Heat shock protein, Hsp90α, transitioning from the closed state to the open state. In Fig. 3(f), two domains of AKT1 kinase coalesce, forming a pocket that did not previously exist.

Taken together, the present model can predict diverse types of conformational changes associated with ligand binding when the ligand-binding pocket is either insufficiently spacious or unformed in the AlphaFold-predicted conformations.

4 Discussion

DynamicBind presents an innovative solution to the challenge of 'dynamic docking' by integrating two traditionally distinct steps—protein conformation generation and ligand pose prediction—into a unified framework. Capable of carrying out substantial conformational changes in protein structures, DynamicBind eliminates the necessity for holo-structures and pre-defined ligand binding sites. These advantages make DynamicBind a powerful tool for a widely range of structure-based drug discovery applications, including virtual screening, discovering cryptic pockets, minimizing side effects of drug candidates, and identifying the pivotal protein targets underlies a disease. Additionally, the ligand-specific protein conformations generated by DynamicBind may offer valuable insights into the influence of ligands on proteins, potentially clarifying structure-function relationships and augmenting our mechanistic understanding.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[2] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[3] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[4] Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pages 2022–07, 2022.

[5] Thomas J Lane. Protein structure prediction has reached the single-structure frontier. *Nature Methods*, pages 1–4, 2023.

[6] Hans Frauenfelder, Stephen G Sligar, and Peter G Wolynes. The energy landscapes and motions of proteins. *Science*, 254(5038):1598–1603, 1991.

[7] Ruth Nussinov, Mingzhen Zhang, Yonglan Liu, and Hyunbum Jang. Alphafold, allosteric, and orthosteric drug discovery: Ways forward. *Drug Discovery Today*, page 103551, 2023.

[8] David D Boehr, Ruth Nussinov, and Peter E Wright. The role of dynamic conformational ensembles in biomolecular recognition. *Nature chemical biology*, 5(11):789–796, 2009.

[9] K Gunasekaran, Buyong Ma, and Ruth Nussinov. Is allostery an intrinsic property of all dynamic proteins? *Proteins: Structure, Function, and Bioinformatics*, 57(3):433–443, 2004.

[10] Maarten L Hekkelman, Ida de Vries, Robbie P Joosten, and Anastassis Perrakis. Alphafill: enriching alphafold models with ligands and cofactors. *Nature Methods*, 20(2):205–213, 2023.

[11] Christoph Gorgulla. Recent developments in structure-based virtual screening approaches. *arXiv preprint arXiv:2211.03208*, 2022.

[12] Richard A Friesner, Jay L Banks, Robert B Murphy, Thomas A Halgren, Jasna J Klicic, Daniel T Mainz, Matthew P Repasky, Eric H Knoll, Mee Shelley, Jason K Perry, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of medicinal chemistry*, 47(7):1739–1749, 2004.

[13] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2): 455–461, 2010.

[14] Valeria Scardino, Juan I Di Filippo, and Claudio N Cavasotto. How good are alphafold models for docking-based virtual screening? *Iscience*, 26(1), 2023.

[15] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *International Conference on Learning Representations (ICLR)*, 2023.

[16] Edward B Miller, Robert B Murphy, Daniel Sindhikara, Kenneth W Borrelli, Matthew J Grisewood, Fabio Ranalli, Steven L Dixon, Steven Jerome, Nicholas A Boyles, Tyler Day, et al. Reliable and accurate solution to the induced fit docking problem for protein–ligand binding. *Journal of Chemical Theory and Computation*, 17(4):2630–2639, 2021.

[17] Pelin Ayaz, Agatha Lyczek, YiTing Paung, Victoria R Mingione, Roxana E Iacob, Parker W de Waal, John R Engen, Markus A Seeliger, Yibing Shan, and David E Shaw. Structural mechanism of a drug-binding process involving a large conformational change of the protein target. *Nature Communications*, 14 (1):1885, 2023.

[18] Xingcheng Lin, Nicholas P Schafer, Wei Lu, Shikai Jin, Xun Chen, Mingchen Chen, José N Onuchic, and Peter G Wolynes. Forging tools for refining predicted protein structures. *Proceedings of the National Academy of Sciences*, 116(19):9400–9409, 2019.

[19] Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv*, pages 2022–12, 2022.

[20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[21] Zhuoran Qiao, Weili Nie, Arash Vahdat, Thomas F Miller III, and Anima Anandkumar. State-specific protein-ligand complex structure prediction with a multi-scale deep generative model, 2023.

[22] Shuya Nakata, Yoshiharu Mori, and Shigenori Tanaka. End-to-end protein–ligand complex structure generation with diffusion-based generative models. *BMC bioinformatics*, 24(1):1–18, 2023.

[23] Bowen Jing, Ezra Erives, Peter Pao-Huang, Gabriele Corso, Bonnie Berger, and Tommi Jaakkola. Eigenfold: Generative protein structure prediction with diffusion models. *arXiv preprint arXiv:2304.02198*, 2023.

[24] Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31(3):405–412, 2015.

[25] Rita Santos, Oleg Ursu, Anna Gaulton, A Patrícia Bento, Ramesh S Donadi, Cristian G Bologa, Anneli Karlsson, Bissan Al-Lazikani, Anne Hersey, Tudor I Oprea, et al. A comprehensive map of molecular drug targets. *Nature reviews Drug discovery*, 16(1):19–34, 2017.

[26] Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. In *International Conference on Machine Learning*, pages 20503–20521. PMLR, 2022.

[27] Wei Lu, Qifeng Wu, Jixian Zhang, Jiahua Rao, Chengtao Li, and Shuangjia Zheng. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in Neural Information Processing Systems*, 2022.

[28] Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022.

[29] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

[30] William Griffith McBride. Thalidomide and congenital abnormalities. *Lancet*, 2(1358):90927–8, 1961.

[31] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.

[32] Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729*, 2022.

[33] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[35] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976.

[36] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21): 2722–2728, 2013.

[37] Stepan S Batsanov. Van der waals radii of elements. *Inorganic materials*, 37(9):871–885, 2001.

[38] Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):1–20, 2021.

[39] Georgi K Kanev, Chris de Graaf, Bart A Westerman, Iwan JP de Esch, and Albert J Kooistra. Klifs: an overhaul after the first 5 years of supporting kinase research. *Nucleic acids research*, 49(D1):D562–D569, 2021.

[40] John W Lampe, Joshua S Alford, P Ann Boriak-Sjodin, Dorothy Brach, Kat Cosmopoulos, Kenneth W Duncan, Sean T Eckley, Megan A Foley, Darren M Harvey, Vinny Motwani, et al. Discovery of a first-in-class inhibitor of the histone methyltransferase setd2 suitable for preclinical studies. *ACS Medicinal Chemistry Letters*, 12(10):1539–1545, 2021.

[41] Joshua S Alford, John W Lampe, Dorothy Brach, Richard Chesworth, Kat Cosmopoulos, Kenneth W Duncan, Sean T Eckley, Jeffrey L Kutok, Alejandra Raimondi, Thomas V Riera, et al. Conformational-design-driven discovery of ezm0414: a selective, potent setd2 inhibitor for clinical studies. *ACS Medicinal Chemistry Letters*, 13(7):1137–1143, 2022.

[42] Mengyao Zhao, Wan-Ping Lee, Erik P Garrison, and Gabor T Marth. Ssw library: an simd smith-waterman c/c++ library for use in genomic applications. *PloS one*, 8(12):e82138, 2013.

[43] Petra Krafcikova, Jan Silhan, Radim Nencka, and Evzen Boura. Structural analysis of the sars-cov-2 methyltransferase complex involved in rna cap creation bound to sinefungin. *Nature communications*, 11 (1):3717, 2020.

[44] Felix Wong, Aarti Krishnan, Erica J Zheng, Hannes Stärk, Abigail L Manson, Ashlee M Earl, Tommi Jaakkola, and James J Collins. Benchmarking alphafold-enabled molecular docking predictions for antibiotic discovery. *Molecular Systems Biology*, 18(9):e11081, 2022.