Modeling protein-small molecule conformational ensembles with ChemNet

Ivan Anishchenko[1,2†], Yakov Kipnis[1,2,4†], Indrek Kalvet[1,2,4†], Guangfeng Zhou[1,2], Rohith Krishna[1,2], Samuel J. Pellock[1,2], Anna Lauko[1,2,3], Gyu Rie Lee[1,2,4], Linna An[1,2], Justas Dauparas[1,2], Frank DiMaio[1,2], David Baker[1,2,4*]

†These authors contributed equally
*To whom correspondence should be addressed: dabaker@uw.edu
1. Department of Biochemistry, University of Washington, Seattle, WA 98105, USA
2. Institute for Protein Design, University of Washington, Seattle, WA 98105, USA
3. Graduate Program in Biological Physics, Structure and Design, University of Washington, Seattle, WA 98105, USA
4. Howard Hughes Medical Institute, University of Washington, Seattle, WA 98105, USA

Abstract

Modeling the conformational heterogeneity of protein-small molecule systems is an outstanding challenge. We reasoned that while residue level descriptions of biomolecules are efficient for de novo structure prediction, for probing heterogeneity of interactions with small molecules in the folded state an entirely atomic level description could have advantages in speed and generality. We developed a graph neural network called ChemNet trained to recapitulate correct atomic positions from partially corrupted input structures from the Cambridge Structural Database and the Protein Data Bank; the nodes of the graph are the atoms in the system. ChemNet accurately generates structures of diverse organic small molecules given knowledge of their atom composition and bonding, and given a description of the larger protein context, and builds up structures of small molecules and protein side chains for protein-small molecule docking. Because ChemNet is rapid and stochastic, ensembles of predictions can be readily generated to map conformational heterogeneity. In enzyme design efforts described here and elsewhere, we find that using ChemNet to assess the accuracy and pre-organization of the designed active sites results in higher success rates and higher activities; we obtain a preorganized retroaldolase with a k_{cat}/K_M of 11000 $M^{-1}min^{-1}$, considerably higher than any pre-deep learning design for this reaction. We anticipate that ChemNet will be widely useful for rapidly generating conformational ensembles of small molecule and small molecule-protein systems, and for designing higher activity preorganized enzymes.

Main Text

Interactions of proteins with nucleic acids, small organic and inorganic molecules, and metals are critical to biological function, but atomistic modeling of such interactions and their conformational heterogeneity remains a challenging problem. Deep learning (DL)-based small molecule docking tools like DiffDock[1] improve on earlier methods such as Glide[2] and GNINA/SMINA, but in the high accuracy regime the difference in performance is not pronounced, and performance also drops substantially on unseen receptors[1,3,4]. DL methods have also been devised to generate small molecule conformations from their chemical structure[5–7]; when trained on the synthetic GEOM-DRUGS dataset[8] of drug-like molecules computed with semi-empirical QM methods, diffusion generative models like DiffDock[1] and Torsional Diffusion[7] show best-in-class performance on the hold-out test set. However, these approaches model specific classes of interaction partners, limiting the ability to model the full spectrum of protein functions, and the input features can differ depending on the type of input molecules, which could limit the ability of the networks to distill general physico-chemical principles. AlphaFold2 (AF2)[9] and RoseTTAFold (RF)[10] enabled atomically accurate structure prediction of proteins and protein-protein complexes using sequences and structures of evolutionary-related proteins as inputs. These methods have recently been extended to modeling the structures of protein-nucleic acid (NA) complexes and more general biomolecular systems[11–13] by supplementing tokens representing residues with full atomic representations of the non-protein components of the system. While the accuracy of AlphaFold3 (AF3) appears to be higher than previous methods, the method is not currently available for general use, and it is unclear whether the method can model conformational heterogeneity.

We reasoned that higher accuracy in modeling structure and conformational heterogeneity could be achieved by an atom centric approach to predicting the structures of small molecules and peptides in isolation and in the context of a protein binding or active site defined at the backbone level. Networks such as AF2, AF3, and RF achieve accurate structure prediction using a primarily residue-level description of biomolecules and multiple sequence information to help infer contacts. At the atomic level close to the native structure, the complexity grows considerably and evolutionary information is less relevant. Hence, we reasoned that an appropriate starting point for an ensemble generation method would not be the amino acid sequence of a protein but rather the coordinates of the overall protein backbone surrounding a binding or catalytic pocket, along with an atomic level description of the bonded geometry of the interacting small molecules and amino acid side chains. Such a network would, by construction, not be useful for structure prediction from sequence, but could be very useful for modeling the conformations of small molecules and constrained peptides both in isolation and in a protein binding site along with the conformations of the interacting sidechains. Since the protein backbone structure is input, the calculations could be faster than AF2, AF3, and RF, enabling rapid binding site refinement and evaluation. We reasoned that a network that generated predictions stochastically could have the advantage of enabling rapid generation of predicted ensembles for modeling systems with a distribution of possible states (which is difficult with AF2, AF3, and RF), and for evaluating the extent of preorganization of designed molecules and functional sites.

Guided by this reasoning we set out to develop a stochastic deep neural network called ChemNet for atomistic modeling of small molecules and small molecule-protein interactions (Fig. 1A). In many applications, a reliable structure of the target protein can be generated by AF2[9] or RF[10], or obtained from the PDB, and location of the putative interacting region on the protein surface is also known; the task is then to dock a small molecule into the region of interest, while adjusting conformations of the protein side chains and the molecule itself. We framed the learning problem as a structure denoising task and trained ChemNet to recapitulate the correct atom positions from partially corrupted input structures provided that all the chemical information about the system being modeled is known from the start. We customized the corruption strategy to the application of interest. In the case of the protein-small molecule docking (Fig. 1A),

the inputs to the network include the protein backbone coordinates, the amino acid sequence with side chain coordinates randomly initialized around the respective C-alpha atoms, and the small molecule chemical structure with atomic positions randomly initialized in the vicinity of the putative binding site.

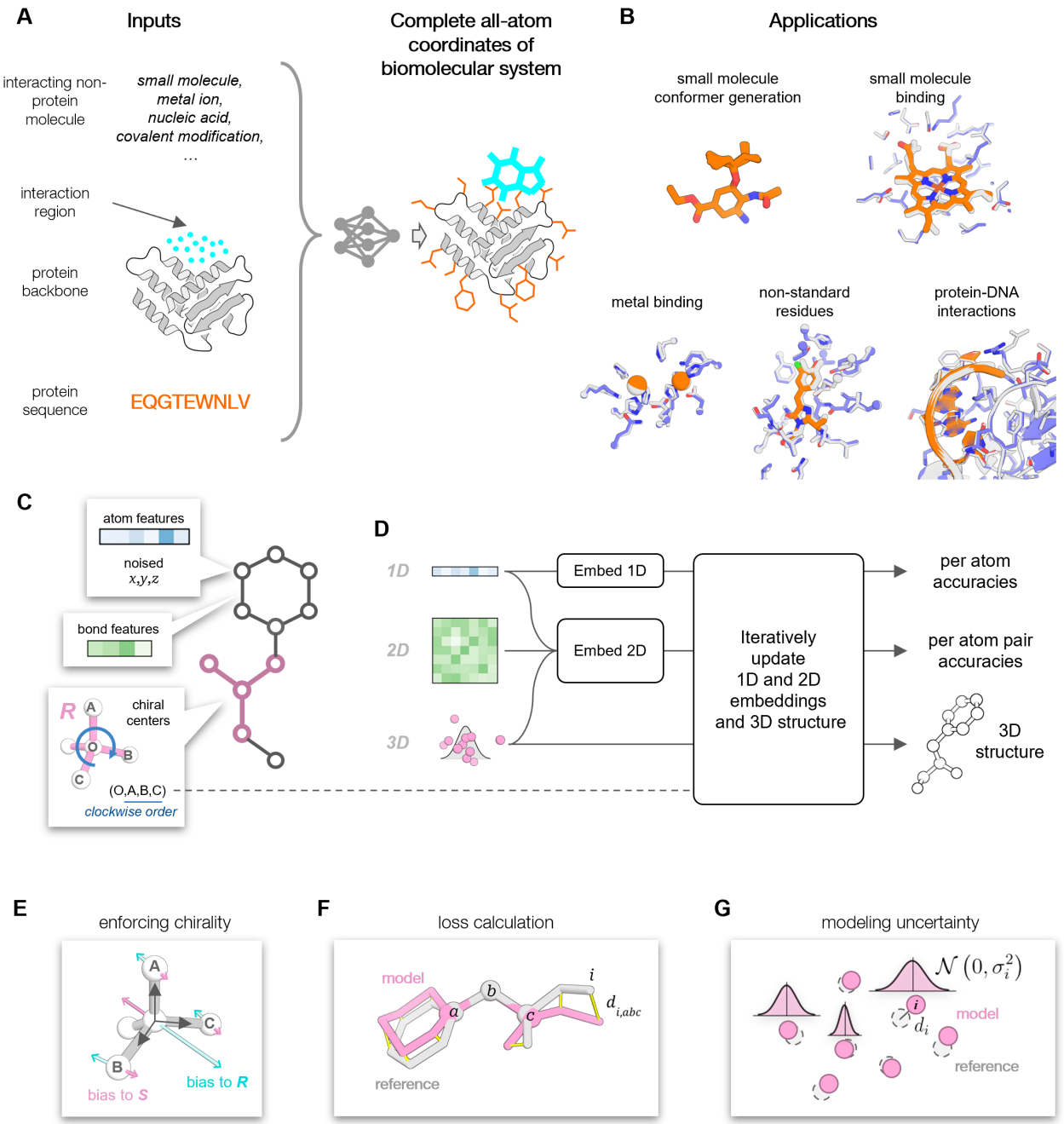

Figure 1. **Overview of ChemNet**. **A**) ChemNet is a denoising neural network which takes at input a partially corrupted protein structure and the chemical structure (but not the coordinates) of any interacting molecules, and predicts the all-atom structure of the complex, as well as the uncertainties in the atom positions in the generated model. **B**) ChemNet can be used for a wide range of tasks including docking of small molecules and metals to a protein target, modeling non-standard residues, and predicting side chain conformations of amino acid residues and nucleotides at the protein-DNA interface. Shown are x-ray structures (in gray) superimposed with ChemNet models (in blue and orange). **C**) At input, the molecular system is represented by an annotated graph where nodes are individual atoms and edges are chemical bonds between atoms. Information about chiral centers is supplied to the network as (O,A,B,C) tuples where

O is the central pyramidal or tetrahedral atom and its neighboring atoms A,B,C are ordered clockwise. **D**) ChemNet is a three-track network that iteratively updates 1D and 2D embeddings and the 3D structure, producing at each iteration a refined atomic structure model and estimating uncertainties in atom placements. **E**) The triple product $V = \vec{e}_A \cdot \vec{e}_B \times \vec{e}_C$ of the three unit vectors pointing from the central chiral atom to its neighbors (gray arrows) is a pseudoscalar that differs in sign for the *R* and *S* configurations: for ideal tetrahedral geometry $V_R = \frac{-4}{3\sqrt{3}}$ and $V_S = \frac{+4}{3\sqrt{3}}$. By comparing V in the non-ideal geometry of the modeled structure to the ideal values V_R or V_S and taking gradients w.r.t. atom coordinates, one gets biasing vectors $\vec{f}_{R/S} = \nabla_{\vec{r}}\left(V - V_{R/S}\right)^2$ showing the directions in which atoms should be moved around in order to recreate the desired configuration. **F**) All-atom $FAPE$ is calculated by aligning the model and the reference structures on every three respective bonded atoms a, b, c and calculating the deviations in atom positions between the aligned structures. $FAPE_{all\ atom}$ is then the mean over all atoms and all superimpositions. Atom-atom distances are clamped at 10Å. **G**) Assuming that uncertainties in atom positions in the modeled structure are normally distributed, we let the dedicated head of the network predict the variances ${\sigma_i}^2$ for every atom in the system to recapitulate actual deviations d_i. These variances are learned during training by maximizing the likelihood $N(d_i; 0, {\sigma_i}^2)$.

At input, the molecular system is converted to a chemical graph with nodes representing individual heavy atoms (hydrogen atoms are not modeled to reduce computation cost) and edges representing chemical bonds between atoms (Fig. 1C). This representation is uniform across molecule types. Each node in the network carries information about the atom type and its 3D coordinates which are initially corrupted. The network is tasked to iteratively denoise the input coordinates and to estimate uncertainties in the atom positions of the output model structure (Fig. 1D). ChemNet has a three-track architecture inspired by RF [10]. After the initial embedding of the 1D and 2D features, they are passed to a block which iteratively updates the embeddings and the 3D structure. In the iteration block, the atom neighbor graph is first constructed: for every atom, the 32 closest neighbors are picked in equal proportions based on both spatial proximity and proximity in the chemical graph. 2D pair features are then projected by a feed-forward adapter layer to edge embeddings, and, together with the 1D features, the atom neighbor graph and the current atomic 3D structure serve as inputs to the SE3-Transformer network[14] which updates the 3D coordinates and the 1D embeddings. Information about the chiral centers is communicated to the network via *type-1* (vector) features (Fig. 1E). Features in the 2D track undergo pair-to-pair updates with bias from structure[15]. Atom and atom pair confidence prediction heads branch off the 1D and 2D tracks respectively, completing the iteration block. The fully trained network consists of eight iteration blocks with shared weights.

ChemNet is trained using a combination of structure and confidence prediction losses applied after every iteration. The primary structure loss is all-atom frame-aligned point error $FAPE_{all\ atom}$, which is an extension of the original $FAPE$ from Ref [9] to arbitrary molecules (Fig. 1F). The confidence of the modeled structures is evaluated on a per atom and per atom pair basis. Atom-atom pair accuracies are estimated using distance signed error approach introduced in Ref [16]. On the per-atom level, we predict all-atom lDDT scores as in AF2 [9]. Additionally, the network predicts deviations in atomic positions σ_i (in Å) in the generated model relative to the reference structure (Fig. 1G). These predicted deviations can then be combined over a subset of atoms of interest (e.g., a small molecule) to give the predicted RMSD $pRMSD = \left(\frac{1}{N}\sum_{i=1}^{N} {\sigma_i}^2\right)^{1/2}$.

Tuning network architecture on experimental structures of small molecules

While small rigid molecules have known three-dimensional structures, larger organic molecules can have unique ground states that are very time consuming to predict with computational methods, hence, accurate prediction of general molecular structure is an important challenge. We explored network architecture and training hyperparameters on crystal structures of chemicals from the Cambridge Structural Database[17]. This also helped ensure that the network architecture can handle diverse chemistry and considerably reduced the architecture exploration time (ChemNet training on the PDB until convergence takes x10

longer). The training task was to predict the small molecule conformation observed in the crystal from its chemical structure and randomly initialized starting coordinates (Fig. 2A). As ChemNet is a denoising network, structurally diverse samples of molecular conformations can be generated by running the network multiple times with different random initialization of the input coordinates. Training of ChemNet on CSD structures was done in two stages: in the first stage, four iteration blocks were used and $FAPE_{all\ atom}$ was the only loss term. In the second stage, the number of iterations was increased to eight and bonded geometry loss terms were added to enhance the local quality of predicted structures (red vs violet on Figs. 2B and C). Reducing the number of iterations to two or replacing the $FAPE$ loss with a combination of coordinate and distance RMSD losses both result in predictions of substantially lower quality (orange and blue on Figs. 2B and C). Performance also degraded when the 2D inputs did not include the bond separation feature (an integer counting the number of covalent bonds between any two atoms in the chemical graph) but only included a binary feature flagging that the two atoms are immediately connected by a bond. Fully trained ChemNet generates the correct 3D structures of molecules as complex as macrocycles with over 50 atoms (Fig. 2D).

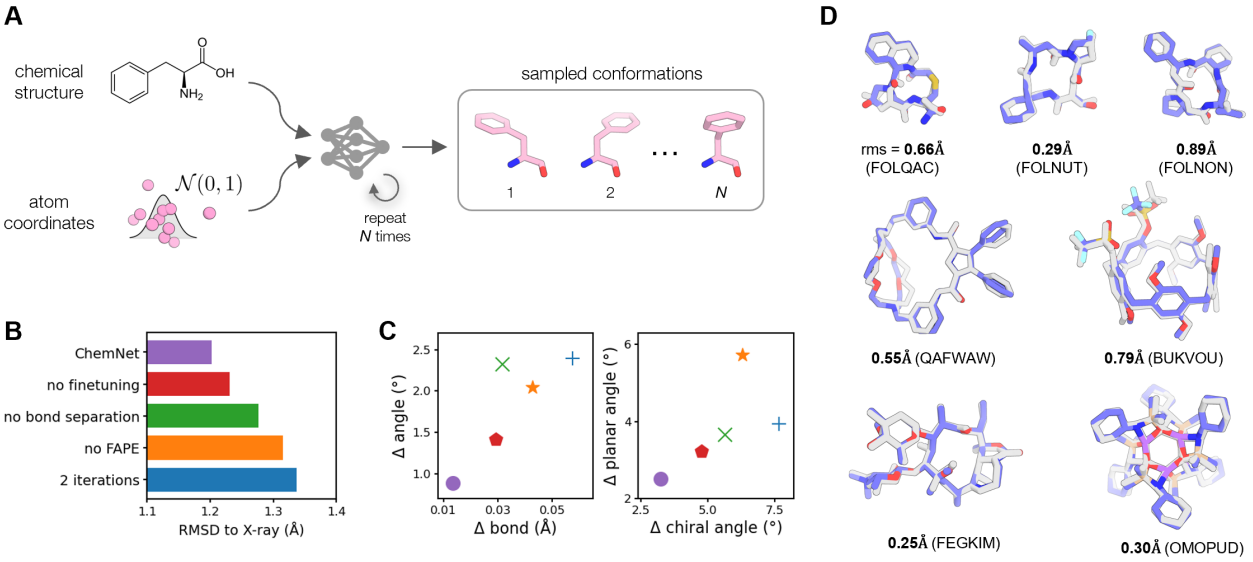

Figure 2. **Modeling complex small molecules**. **A)** ChemNet predicts the 3D structure of a small molecule from its chemical structure and randomly initialized atomic coordinates. Running ChemNet multiple times with different random initializations of atom positions yields a diverse set of molecular conformations. **B)** Ablating ChemNet features negatively affects performance as measured by the root-mean-square deviation in atom coordinates between the model and the reference x-ray structure after the two are optimally superimposed. **C)** Quality of the local geometry in models produced by the five networks from panel B. Shown are mean absolute errors in bonds, bonded, planar, and chiral angles computed from the model and reference structures. Validation set of 7,117 small molecules from the CSD was used for panels B and C. **D)** Examples of seven macrocyclic molecules for which ChemNet generates the correct structure. Shown are the best RMSD models (in blue; out of 1000 generated) overlaid with experimental crystal structures (gray). CCDC database identifiers shown in parentheses.

Modeling protein-small molecule interactions with ChemNet

To train ChemNet on the structures from the PDB (proteins plus small molecules, rather than small molecules alone, as in the case of the CSD) we parsed them into chemical graphs (Fig. 1C) using the Chemical Component Dictionary[18] that provides detailed chemical description for all residue and small molecule components found in the PDB. We reasoned that although some molecules in the PDB may be non-biological (e.g., solvents) and their interactions with the rest of the structure may be non-specific, these interactions could still carry information about physico-chemical preferences at molecular interfaces and

thus be informative for the network training; only water molecules were excluded. Training and validation sets (112,828 and 7,090 examples, respectively) were compiled from higher resolution (<2.5Å) structures deposited to the PDB before January 12, 2023. During training, the input structures were cropped to at most 600 heavy atoms, and centered around a randomly selected atom. The non-backbone-atom coordinates of all individual connected molecular components in the crop were collapsed to their connected backbone atom (if residue or nucleic acid), or to a randomly selected atom (ligands), and corrupted with Gaussian noise (σ = 1.5Å). From these starting points, ChemNet was trained to recapitulate the atomic coordinates of all atoms in the cropped regions. For a detailed description of the training procedure, see Supplementary Methods.

We used ChemNet to generate ensembles of small molecule poses in the pocket of the target protein (Fig. 3A) by repeated runs with different random initialization of the input coordinates. Analyzing such ensembles reveals that ChemNet is not very sensitive to the initial placement of the ligand: multiple different starting positions yield near native predictions (Fig. 3B), and these positions cover the entire space sampled at input (Fig. 3D). We also observed that the predicted $pRMSD$ score calculated over the ligand atoms allows selection of more accurate models from the sampled pool (Fig. 3E), with the best scoring models closely matching the experimental structure (ligand RMSD = 0.53Å and 0.77Å for heme[19] and cortisol[20] respectively, Fig. 3C).

To assess ChemNet performance in a more realistic scenario of docking against non-native protein conformations, we used a standard test set of 65 drug targets [21]. Each of these targets has a co-crystal structure with a drug molecule, as well as a number of structures either without the drug or with a different small molecule co-crystallized, totaling to 1112 non-native structures over the 65 targets. The docking results highlight the importance of both sampling and scoring aspects of the network (Fig. 3F): the success rate of generating and selecting a near-native conformation of the docked small molecule increases if more models are generated. Among the three confidence prediction metrics tested, the $pRMSD$ score shows the best power of selecting near-native conformations (Fig. 3G, top 3 bars), with higher docking success rates than Vina, GOLD, and GalaxyDock (Fig. 3G, bars in the middle). Compared to the best performing Rosetta GALigandDock method,[22] ChemNet performance is higher in the lower accuracy regime (% of complexes with ligand RMSD < 2Å, 82.4% vs 73.6%), but is behind in the high-accuracy regime (RMSD < 1Å, 41.8% vs 51.6%). ChemNet performance is still notable however, because, unlike the other methods, the network was not specifically trained on the non-native protein-small molecule docking task. Additionally, ChemNet recreates both the conformations of the small molecule and the protein side chains from scratch, while the other methods tested rely largely on the input protein coordinates. Improved performance can be obtained by combining ChemNet with physics-based methods: minimizing the docks generated by the network using the generalized Rosetta force field and estimating binding free energies of the minimized structures gives +7.3% increase in docking success rate at <1Å and no notable change at <2Å; likely due to improved sampling.

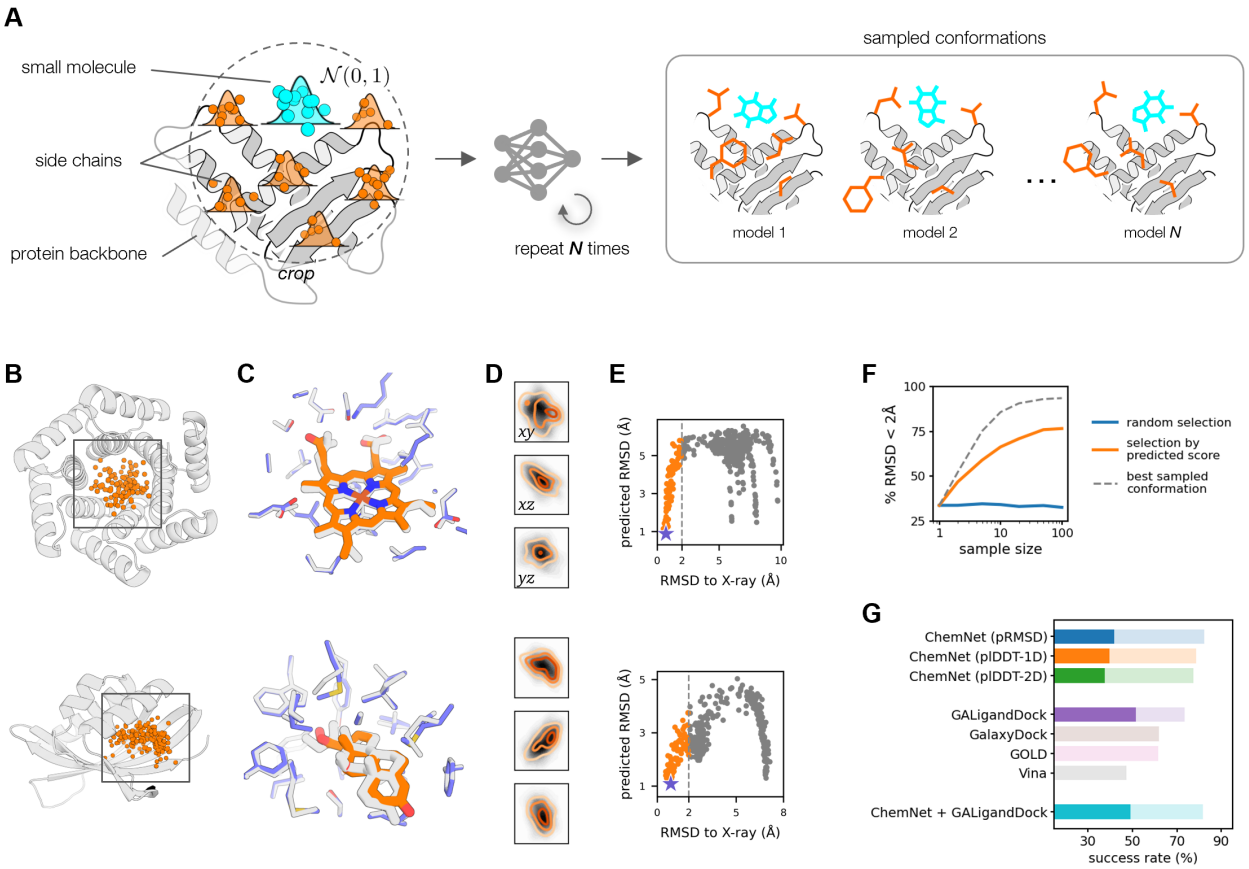

Figure 3. **Modeling protein- small molecule interactions with ChemNet**. **A)** Starting with the protein backbone and the coordinates of the small molecule randomly initialized in the vicinity of the binding site (cyan dots), ChemNet predicts protein side chain conformations (which are initially randomized around the respective C-alpha atoms, orange dots) and the structure and the placement of the small molecule relative to the target protein. By repeating this process multiple times ChemNet generates a structurally diverse set of docks. **B,D)** Sampled starting positions of the small molecule for the two de novo designed small molecule binders. Panel **D** shows projections of the sampled starting points onto the three coordinate planes. Shades of gray show density of all sampled positions; orange contours indicate density of points which resulted in docks with ligand RMSD < 2Å. The two densities largely overlap, suggesting that ChemNet is not very sensitive to the initial placement of the ligand in the binding pocket. **C)** Poses with the lowest $pRMSD$ scores (blue/orange) closely match the experimental structure (gray). **E)** Models with lower $pRMSD$ scores are funneled towards the native conformation demonstrating the discriminative power of the predicted score. **F)** Increasing the sample size and rescoring ChemNet models by the $pRMSD$ score (orange curve) greatly increases the chance of picking a better model compared to a random selection (blue curve). Scoring by $pRMSD$ is however still not optimal (orange vs gray dashed curve). **G)** Among the three accuracy scores predicted by ChemNet, $pRMSD$ shows best performance (the three bars at the top). Bright and pale bars show docking success rates in the Astex non-native set for ligand RMSD < 1Å and 2Å respectively. The four bars in the middle show performance of the widely used docking tools Vina, GOLD, GalaxyDock, and Rosetta GALigandDock. Sampling docks with ChemNet and minimizing and rescoring them using GALigandDock increases docking success rate by 7.3% (ligand RMSD < 1Å) over ChemNet alone.

Assessment of enzyme active site design accuracy and pre-organization

A critical challenge in de novo enzyme design is to come up with amino acid sequences that not only encode the target backbone structure but also position the catalytic sidechains and the reaction substrate such that the key sidechain functional groups make the hydrogen bonding and electrostatic interactions with the transition state and each other that are essential to catalysis. As 0.5 Å deviations in distance and 30 degree deviations in angle can have large effects on hydrogen bonding energies, the level of accuracy in positioning required is quite high. Furthermore, the designed active site should be pre-organized in the absence of substrate with the catalytic sidechains largely held in place by intra-protein interactions to reduce the entropic cost of binding and ensure that all catalytic groups are properly positioned; such pre-organization is a notable feature of natural enzyme active sites. The most general way to evaluate such pre-organization is molecular dynamics simulations[23,24], but as sidechain reconfiguration can occur on the microsecond time scale, such simulations are computationally expensive and cannot be applied to large numbers of designs. Approximate discrete methods such as the Rosetta Rotamer Boltzmann method[25] estimate pre-organization at the individual sidechain level, but are limited by the rotamer approximation and the inability to model coupled and concerted movements of side chains and small molecules. We reasoned that ChemNet could enable assessment of the accuracy and pre-organization of designed active sites by generating ensembles of models of small molecules and interacting sidechain placements much more rapidly than molecular dynamics simulations and without the limitations of the Rotamer Boltzmann method.

We applied ChemNet to the RA95 series of previously designed retro-aldolase (Fig. 4A) enzymes, and directed evolution generated improved versions of these, for which high resolution crystal structures have been determined[26–28]. As illustrated in Figs. 4B and 4C, ChemNet generated ensembles are highly varied for the lower activity initial computational designs, indicating a lack of pre-organization, and become increasingly well ordered for the more active evolved versions. This suggests that lack of pre-organization was a major shortcoming of early enzyme design efforts (similar conclusions were reached using considerably more expensive MD simulations[29]), and that ChemNet provides a rapid means to assess this, and thus guide enzyme design efforts.

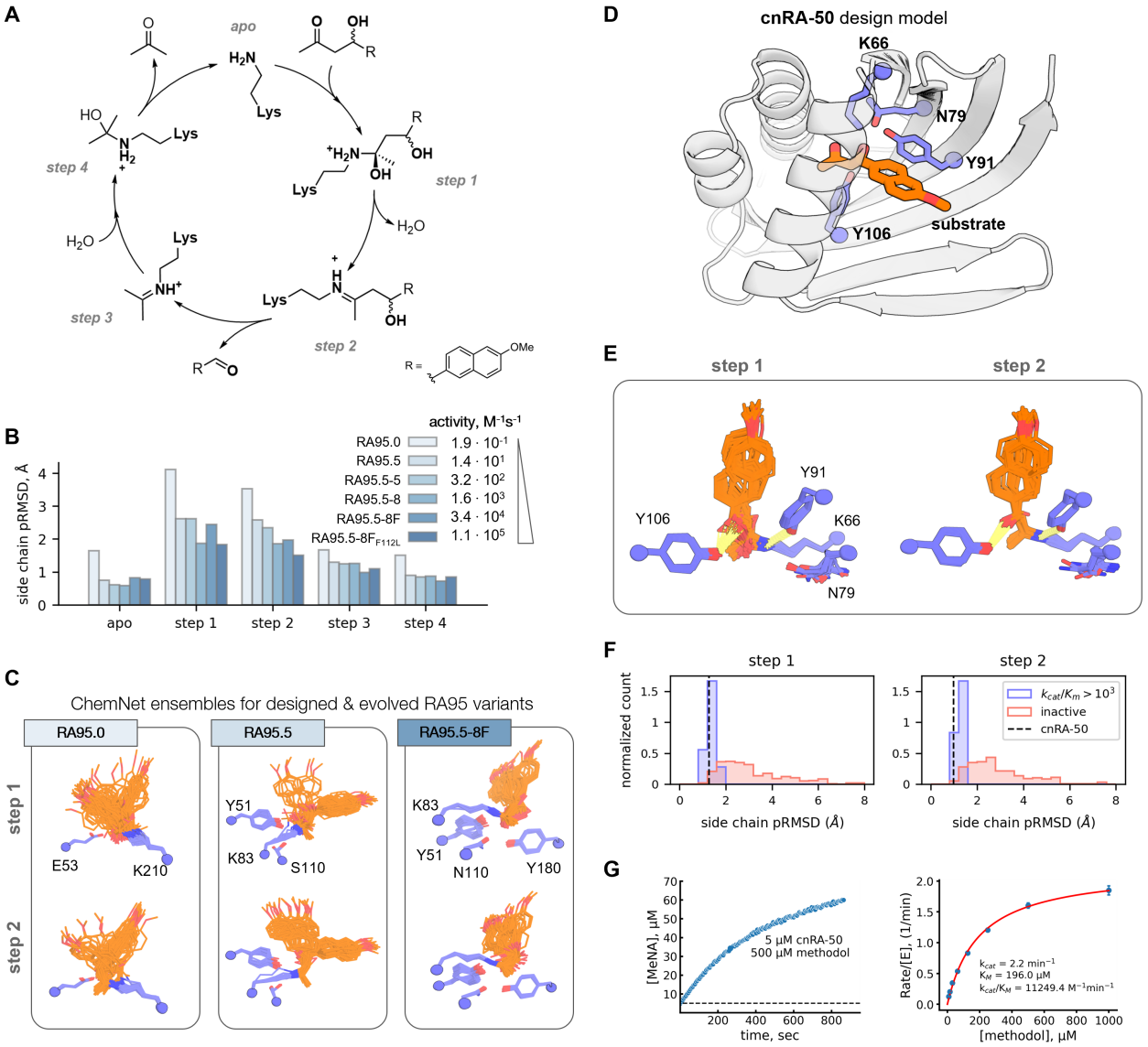

Figure 4. **Selecting designs with preorganized catalytic residues increases activity**. **A)** Mechanism of the retro-aldol reaction with the key intermediates shown. **B)** Five intermediate states of the RA reaction with methodol were modeled by ChemNet for the RA95 series of designs to probe preorganization of the active site lysine and its modifications. **C)** Examples of ChemNet ensembles for active sites of the three previously published retro-aldolases with increasing activity show higher degree of preorganization in steps 1 and 2 along the reaction pathway. **D)** Design model of the most active variant, cnRA-50, with the catalytic tetrad highlighted in blue and substrate in orange. **E)** ChemNet generated ensembles of lysine-methodol conjugates of catalytic steps 1 and 2 of the active site of cnRA-50. **F)** ChemNet active site pre-organization ensemble metrics (pRMSD of conjugated lysine atoms, except backbone) in steps 1 and 2 enrich for selecting more active de novo designed retro-aldolase enzymes. **G)** Retro-aldol reaction catalyzed by cnRA-50: time-course following the formation of 6-methoxy-2-naphthaldehyde (left), and Michaelis-Menten plot (right) obtained from initial velocities.

Additionally, we tested the use of ChemNet prospectively to guide protein design by making a new round of designed retro-aldolases using deep learning generated NTF2-like fold[20,30] that previously yielded high activity de novo luciferases[31]. We started from an active site description based on the lysine centered catalytic tetrad found in the most active of the evolved retroaldolases[28] (Fig. S3). This active site contains an intricate network of hydrogen bonds keeping the catalytic residues in place (Fig. 4D); we reasoned that

ChemNet might be helpful to overcome this difficulty by assisting in identification of pre-organized active sites. Designs were generated using RosettaMatch[32] and LigandMPNN[33]; full details of the design strategy are provided in the supplemental methods. The activities of the 320 designs were first evaluated in an in vitro transcription translation system (IVTT), enabling rapid identification of promising variants without the need for protein purification. The most active of these were expressed and purified in *E. coli*, and their k_{cat}/K_m values determined by measuring activity as a function of substrate concentration. We then examined how ChemNet preorganization correlated with activity over all of the tested designs, focusing on the conformational flexibility of the catalytic lysine conjugated to reaction intermediates in the ensembles at each step in the reaction pathway (Fig. 4E). We found that the designs with the highest k_{cat}/K_m values were more preorganized than the low activity designs from the initial IVTT screen (Fig. 4F and Fig. S7). The most active of the ChemNet preorganized designs, cnRA-50, had a k_{cat}/K_m of 11,000 $M^{-1}min^{-1}$ (Fig. 4G), much higher than earlier computational designs prior to directed evolution, and comparable to recent designs made using RFdiffusion and proteinMPNN[34].

Conclusions

ChemNet provides a versatile and rapid approach for generating conformational ensembles of molecules both in isolation and in the context of a protein binding site given the sequence of the protein and the positions of the backbone atoms. Unlike AF3, RoseTTAFold All-Atom, and other protein structure prediction methods, ChemNet does not predict protein backbone structure—the benefit is that calculations are considerably more rapid, enabling stochastic generation of conformational ensembles. The use of a consistent atom level representation for all interactions readily allows for extension beyond biomolecules to macrocycles and other complex small molecules. The ability to rapidly generate conformational ensembles for protein-small molecule assemblies should be of considerable utility for computational enzyme design and small molecule binder design efforts: the accuracy with which the intended active sites are recapitulated and the extent of pre-organization of the key catalytic/ interacting sidechain functional groups can be readily assessed. ChemNet assessment of active site accuracy and pre-organization considerably improves discovery success rates for multi-step serine hydrolases[35], and we describe here the design of retroaldolases with much higher activities than pre-deep learning designs for this reaction, prior to experimental optimization. We anticipate that ChemNet based ensemble generation will be broadly useful for modeling the structures of complex non-protein molecules both in isolation and in a protein context, and for evaluating enzyme and protein-small molecule binder designs quite generally.

Acknowledgments:

We thank Luki Goldschmidt and Kandise VanWormer for maintaining the computational and wet lab resources at the Institute for Protein Design. We thank Dr. Declan Evans, and Dr. Florence Hardy for helpful conversations during the development of the method, and Madison A. Kennedy for reading and editing earlier drafts of the manuscript.

Funding:

We thank Microsoft for generous donation of Azure Compute Credits, and Perlmutter grant NERSC award BER-ERCAP0022018 for access to the Perlmutter high performance computing resources. This work was supported by a gift from Microsoft (R.K., D.B., I.A., J.D.). This work was funded by The Howard Hughes Medical Institute (D.B, I.K, G.R.L.), the Schmidt Futures program (F.D.), the Open Philanthropy Project Improving Protein Design Fund (I.K, G.R.L., Y.K., S.J.P., J.D.), the Audacious Project at the Institute for

Protein Design (L.A., A.L.), the Washington Research Foundation's Innovation Fellows Program (G.R.L.), Postdoctoral Fellowship Award (S.J.P.), and Translational Research Fund (L.A.), Defense Threat Reduction Agency HDTRA1-19-1-0003 (S.J.P., A.L.) and HDTRA1-21-1-0007 (G.Z., I.A.), the National Institute of Allergy and Infectious Diseases, National Institutes of Health (NIH), Department of Health and Human Services, under Contract No.: 75N93022C00036 (I.A.), NIH and/or the National Institute of General Medical Sciences (NIGMS) Award (T32GM008268) (A.L.), and the National Science Foundation CHE-2226466 (F.D.).

Author Contributions:

Designed the research: I.A. and D.B. Developed ChemNet architecture and training regimen: I.A. Evaluated ChemNet on different prediction tasks: I.A., Y.K., G.Z., S.J.P., A.L., I.K., L.A., G.R.L. Generated and tested retroaldolase designs: Y.K. Contributed code and ideas: I.A., I.K., R.K, J.D. Offered supervision throughout the project: D.B., F.D. Wrote the manuscript: I.A., Y.K, I.K., and D.B. All authors read and contributed to the manuscript.

Competing interests:

A provisional patent (application number 63/535,404) covering ChemNet network presented in this paper has been filed by the University of Washington. D.B., I.A., G.Z., R.K, F.D. are inventors on this patent. D.B. is a cofounder and shareholder of Vilya, an early-stage biotechnology company that has licensed the provisional patent.

Data and Materials Availability:

Code and neural network weights will be made publicly available on GitHub. All other data is available in main text or Supplemental Materials.

References

1. Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. *arXiv [q-bio.BM]* (2022).
2. Halgren, T. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **47**, 1750–1759 (2004).
3. Stärk, H., Ganea, O.-E., Pattanaik, L., Barzilay, R. & Jaakkola, T. EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction. *arXiv [q-bio.BM]* (2022).
4. Lu, W. *et al.* TANKBind: Trigonometry-Aware Neural NetworKs for drug-protein binding structure prediction. *bioRxiv* 7236–7249 (2022) doi:10.1101/2022.06.06.495043.
5. Ganea, O. *et al.* Geomol: Torsional geometric generation of molecular 3d conformer ensembles. *Adv. Neural Inf. Process. Syst.* **34**, 13757–13769 (2021).
6. Xu, M. *et al.* GeoDiff: a Geometric Diffusion Model for Molecular Conformation Generation. *arXiv [cs.LG]* (2022).
7. Jing, B., Corso, G., Chang, J., Barzilay, R. & Jaakkola, T. Torsional diffusion for molecular conformer generation. *arXiv [physics.chem-ph]* 24240–24253 (2022).
8. Axelrod, S. & Gómez-Bombarelli, R. GEOM, energy-annotated molecular conformations for

property prediction and molecular generation. *Sci Data* **9**, 185 (2022).
9. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
10. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
11. Krishna, R. *et al.* Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024).
12. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
13. Qiao, Z., Nie, W., Vahdat, A., Miller, T. F., III & Anandkumar, A. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nat. Mach. Intell.* **6**, 195–208 (2024).
14. Fuchs, F., Worrall, D., Fischer, V. & Welling, M. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Adv. Neural Inf. Process. Syst.* **33**, 1970–1981 (2020).
15. Baek, M. *et al.* Efficient and accurate prediction of protein structure using RoseTTAFold2. *bioRxiv* 2023.05.24.542179 (2023) doi:10.1101/2023.05.24.542179.
16. Hiranuma, N. *et al.* Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat. Commun.* **12**, 1340 (2021).
17. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr B Struct Sci Cryst Eng Mater* **72**, 171–179 (2016).
18. Westbrook, J. D. *et al.* The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics* **31**, 1274–1278 (2015).
19. Kalvet, I. *et al.* Design of Heme Enzymes with a Tunable Substrate Binding Pocket Adjacent to an Open Metal Coordination Site. *J. Am. Chem. Soc.* **145**, 14307–14315 (2023).
20. Lee, G. R. *et al.* Small-molecule binding and sensing with a designed protein family. *bioRxiv* 2023.11.01.565201 (2023) doi:10.1101/2023.11.01.565201.
21. Verdonk, M. L., Mortenson, P. N., Hall, R. J., Hartshorn, M. J. & Murray, C. W. Protein-ligand docking against non-native protein conformers. *J. Chem. Inf. Model.* **48**, 2214–2225 (2008).
22. Park, H., Zhou, G., Baek, M., Baker, D. & DiMaio, F. Force Field Optimization Guided by Small Molecule Crystal Lattice Data Enables Consistent Sub-Angstrom Protein-Ligand Docking. *J. Chem. Theory Comput.* **17**, 2000–2010 (2021).
23. Rakotoharisoa, R. V. *et al.* Design of efficient artificial enzymes using crystallographically enhanced conformational sampling. *J. Am. Chem. Soc.* **146**, 10001–10013 (2024).
24. Broom, A. *et al.* Ensemble-based enzyme design can recapitulate the effects of laboratory directed evolution in silico. *Nat. Commun.* **11**, 4808 (2020).
25. Fleishman, S. J., Khare, S. D., Koga, N. & Baker, D. Restricted sidechain plasticity in the structures of native proteins and complexes. *Protein Sci.* **20**, 753–757 (2011).
26. Althoff, E. A. *et al.* Robust design and optimization of retroaldol enzymes. *Protein Sci.* **21**, 717–726 (2012).
27. Giger, L. *et al.* Evolution of a designed retro-aldolase leads to complete active site remodeling. *Nat. Chem. Biol.* **9**, 494–498 (2013).
28. Obexer, R. *et al.* Emergence of a catalytic tetrad during evolution of a highly active artificial aldolase. *Nat. Chem.* **9**, 50–56 (2017).

29. Romero-Rivera, A., Garcia-Borràs, M. & Osuna, S. Role of conformational dynamics in the evolution of retro-aldolase activity. *ACS Catal.* **7**, 8524–8532 (2017).
30. Marcos, E. *et al.* Principles for designing proteins with cavities formed by curved β sheets. *Science* **355**, 201–206 (2017).
31. Yeh, A. H.-W. *et al.* De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).
32. Weitzner, B. D., Kipnis, Y., Daniel, A. G., Hilvert, D. & Baker, D. A computational method for design of connected catalytic networks in proteins. *Protein Sci.* **28**, 2036–2041 (2019).
33. Dauparas, J. *et al.* Atomic context-conditioned protein sequence design using LigandMPNN. *bioRxiv* 2023.12.22.573103 (2023) doi:10.1101/2023.12.22.573103.
34. Braun, M. *et al.* Computational design of highly active de novo enzymes. *bioRxiv* 2024.08.02.606416 (2024) doi:10.1101/2024.08.02.606416.
35. Lauko, A. *et al.* Computational design of serine hydrolases. *bioRxiv* 2024.08.29.610411 (2024) doi:10.1101/2024.08.29.610411.