Protein Ensemble Generation through Variational Autoencoder Latent Space Sampling

Sanaa Mansoor[1,2,3], Minkyung Baek[1,2,4], Hahnbeom Park[1,2,5], Gyu Rie Lee[1,2], David Baker[1,2,6]

1. Department of Biochemistry, University of Washington, Seattle, WA 98195, USA.
2. Institute for Protein Design, University of Washington, Seattle, WA 98195, USA.
3. Molecular Engineering Graduate Program, University of Washington, WA 98195, USA.
4. School of Biological Sciences, Seoul National University, Seoul, 08826, Republic of Korea.
5. Brain Science Institute, Korea Institute of Science and Technology, Seoul, 02792, Republic of Korea
6. Howard Hughes Medical Institute, University ofWashington, Seattle, WA 98195, USA.

Abstract

Mapping the ensemble of protein conformations that contribute to function and can be targeted by small molecule drugs remains an outstanding challenge. Here we explore the use of soft-introspective variational autoencoders for reducing the challenge of dimensionality in the protein structure ensemble generation problem. We convert high-dimensional protein structural data into a continuous, low-dimensional representation, carry out search in this space guided by a structure quality metric, then use RoseTTAFold to generate 3D structures. We use this approach to generate ensembles for the cancer relevant protein K-Ras, training the VAE on a subset of the available K-Ras crystal structures and MD simulation snapshots, and assessing the extent of sampling close to crystal structures withheld from training. We find that our latent space sampling procedure rapidly generates ensembles with high structural quality and is able to sample within 1 angstrom of held out crystal structures, with a consistency higher than MD simulation or AlphaFold2 prediction. The sampled structures sufficiently recapitulate the cryptic pockets in the held-out K-Ras structures to allow for small molecule docking.

Main Text

A major challenge in drug discovery is identifying cryptic binding pockets that can be targeted by small molecule drugs (Beglov et al., 2018; Vajda et al., 2018; Vijayan et al., 2015). Despite considerable advances in single state native protein structure prediction with AlphaFold and RoseTTAFold in the past several years, generating plausible ensembles of structures that can be populated upon binding a small molecule, or during protein function, remains an outstanding problem– AlphaFold and RoseTTAFold generate single structures, rather than ensembles. Molecular dynamics (MD) trajectories generate protein ensembles by simulating protein motion around the native structure, and are often used to generate ensembles prior to small molecule docking calculations, but often fail to identify cryptic ligand binding pockets not present in the unbound structure (Cimermancic et al., 2016; Beglov et al., 2018; Vajda et al., 2018; Vijayan et al., 2015) or require very long and hence highly compute-intensive simulations (typically sub-to-several microsecond level) (Kimura et al., 2017; Meller et al., 2023; Sun et al., 2020). Other classical approaches have been used to sample protein conformers through Rosetta (Larson et al., 2002), and loop sampling using kinematic closure (KIC) (Mandell et al., 2009), but have not sampled the types of conformational changes involved in cryptic pocket formation. On the deep

learning side, variational autoencoders which project complex data into a smaller dimension latent space have been used to generate alternative backbones for general protein design tasks such as de novo design of 64 residue backbones (Anand et al., 2018), graph-based protein design (Ingraham et al., 2019) and Ig-fold modeling (Eguchi et al. 2020). VAEs have been used previously to sample the conformational space of proteins, but have required visual inspection of the trained latent space to sample (Tian et al., 2021), or have focused on mapping correlative fluctuations in extensive MD simulations of both the apo and holo state of a target protein (Tsuchiya et al., 2019).

We reasoned that sampling within the latent space of variational autoencoders could provide a solution to the ensemble generation problem for a specific protein sequence. Unlike most previous VAE approaches, which have trained on many different proteins, the challenge of a protein specific VAE is that there is limited training data. We reasoned this limitation could be overcome by supplementing available crystal structures of the protein of interest in alternative conformations with snapshots from short MD trajectories started from each of these structures. For exploring this approach, we chose the critical cancer target K-Ras as a model system due to its considerable therapeutic importance and the many available structures (Pantsar et al. 2019; Liete et al. 2022).

We began by exploring different VAE architectures, training on ensembles of MD simulations from alternate crystals forms of K-Ras (full details in the Methods section), and evaluating the quality of 3D reconstruction following encoding and decoding. For encoding 3D structural information, we chose to use the two-dimensional RoseTTAFold (Baek et al., 2021) template features. The reconstructed template features were then used as input template features for 3D structure generation with RoseTTAFold, along with the amino acid sequence. We evaluate the accuracy of reconstruction by computing the RMSD between the input and output 3D coordinates. The RMSD loss was calculated based on the 2D template features of the input and output 3D structures. We generate new samples by guided exploration in the latent space, followed by 3D coordinate generation with RF (RoseTTAFold, Figure 1).

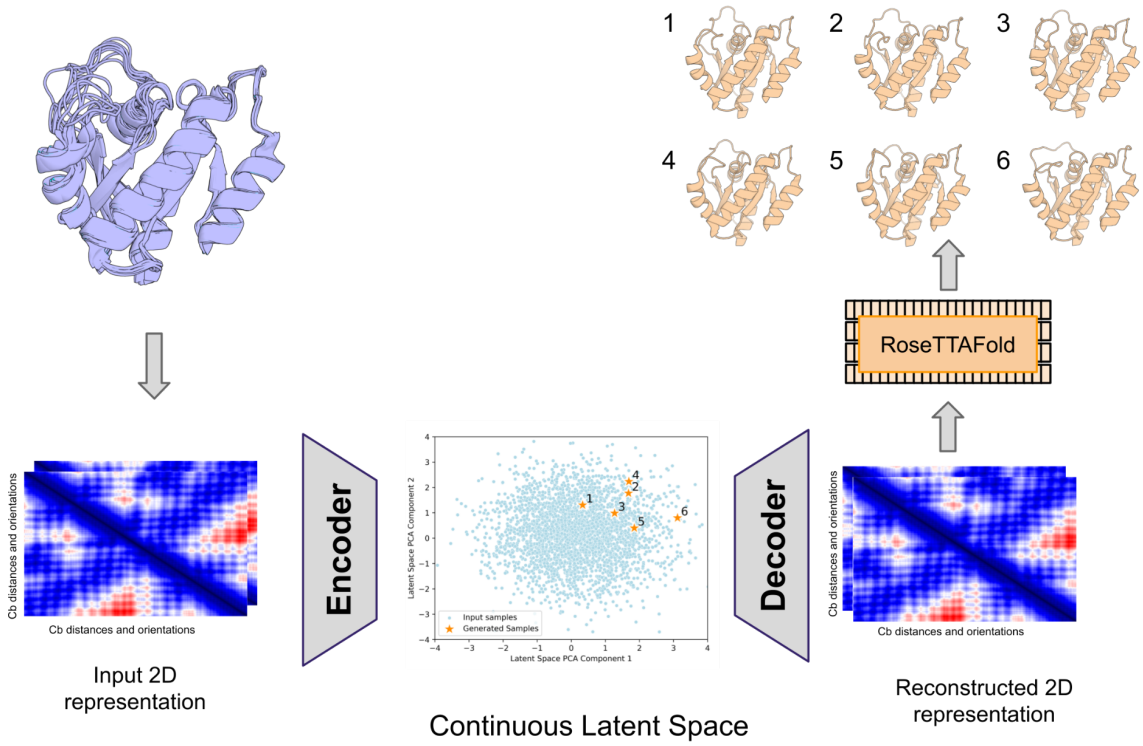

Figure 1. VAE based ensemble generation approach. 3D coordinates from crystal structures and MD simulations are converted to RoseTTAFold 2D template features (Baek et al., 2021). The decoded template features are converted to 3D structures through RoseTTAFold, which is also given the amino acid sequence. Ensembles are generated by sampling in the latent space followed by decoding and RF structure generation.

The reconstruction accuracy of crystal structures not included in the training set provides a rough lower bound on the accuracy with which our approach can recapitulate conformations of interest. For each available K-Ras structure, we trained a VAE leaving out this structure and others within 1A RMSD, and evaluated the accuracy of reconstruction following RoseTTAFold (Baek et al., 2021) 3D coordinate generation. We obtained best results with the soft-introspective VAE architecture (Figure S1), and the accuracy of reconstruction plateaued at ~256 latent space dimensions (Figure S2) For most of the targets (13/20), the reconstructed target from the VAE was of sub-angstrom accuracy (RMSD < 1A); for comparison only 2/20 AF2 structure predictions were of sub-angstrom accuracy (Figure 2).

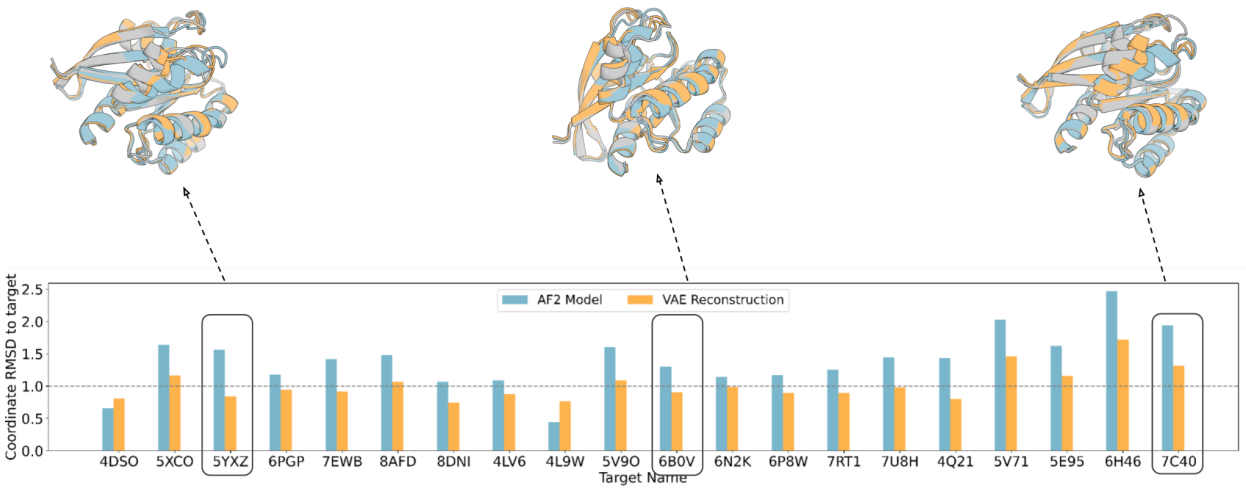

Figure 2. VAE Structure reconstruction accuracy. Coordinate RMSD of the closest AF2 predicted model and the reconstructed model from the VAE decoded template features generated using RoseTTAFold. Structural superimpositions for 3 targets are highlighted on top with the target crystal in gray, the AF2 prediction in blue and the VAE reconstruction in orange.

We next explored the possibility of generating plausible K-Ras ensembles by sampling in the latent space of the trained VAEs. To help ensure that the sampled structures remained broadly consistent with the sequence and were physically plausible, we guided sampling by the consistency to the AF2 predicted distance distribution for the amino acid sequence. Samples were generated from a normal distribution with a mean of 0 and variance of 1, decoded into the corresponding Cb distance map, the CCE to the AF2 predicted distogram for the sequence was computed, and local optimization in the latent space was carried out through gradient descent on the CCE value, limiting the total (latent space) distance traversed from the starting point to prevent convergence. Following decoding and RF structure generation, samples were evaluated using coordinate RMSD to the target crystal on either the overall

structure recapitulation and cryptic pocket environment reconstruction (defined as the residues within 5 angstroms of the ligand binding pocket).

Using this VAE guided sampling approach, we generated K-Ras structure ensembles, again holding out individual K-Ras crystal structures and MD simulation snapshots derived from them, along with other K-Ras crystal structures (and MD snapshots) within 1 angstrom RMSD. We evaluated these ensembles by determining how closely they sampled the held out structures. An advantage of our approach is that ensembles can be generated quite rapidly (compared to MD simulations, for example), and the closest RMSD to the held out structures of course decreases with increasing number of samples (Figure S3). We found that ensembles of 3000 structures sampled more closely to the held out structures than the closest training set crystal structure, training set MD simulation snapshot and the closest AF2 model for most targets (Figure 3).

For small molecule docking calculations, the sampling of alternative ligand binding pocket geometries is particularly important. Comparison of the RMSD over the ligand binding pocket residues between the closest sampled conformation in the generated ensembles and the held out structures showed that the ensembles sample closer than the closest training MD snapshot or crystal structure in most cases (Figure 4). Structural superimpositions show that the generated samples do not clash with the superimposed ligand from the target structure, highlighted in orange, and therefore can be docked without hindrance, whereas for the closest train crystal and the closest AF2 model, there are significant clashes (Figure 4).

We used the physically based GA-ligand docking method to dock ligands onto all the models generated from the VAE, the training examples and the AF2 models. Consistent with the above observations, the RMSD over the ligand atoms was consistently lower for the ensemble generated samples than the AF2 predictions, and lower in most cases than the docks to the MD ensembles (Figure 5).

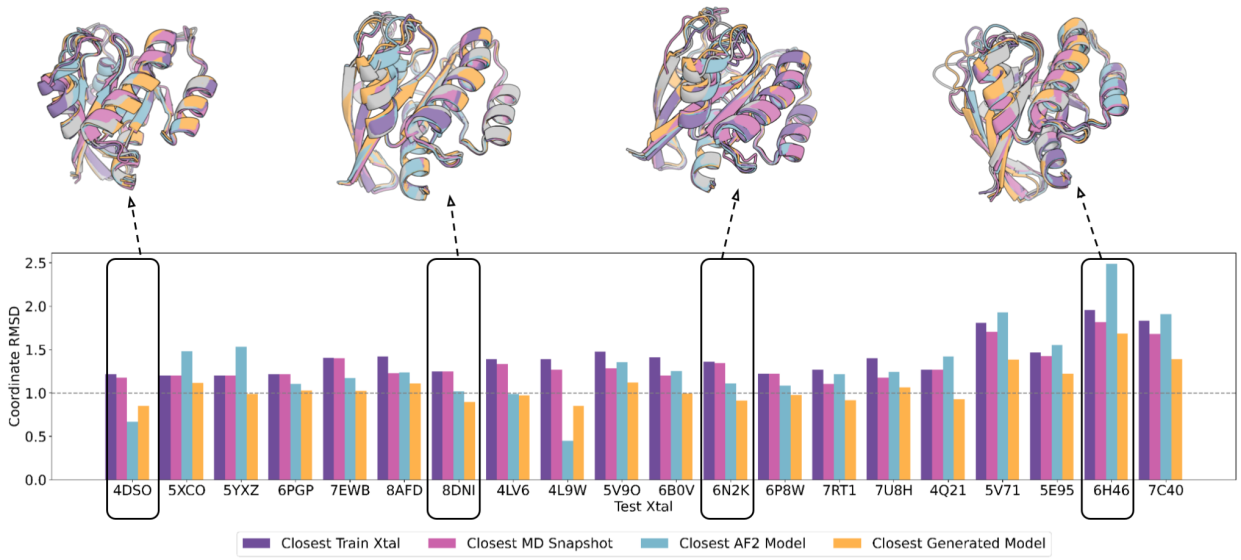

Figure 3. The VAE enables sampling closer to held out K-Ras crystal structures than MD or AlphaFold generated structures. For each test crystal structure (name below bars), a VAE was trained using MD simulation data from all crystal structures with greater than 1A RMSD, and used to generate a structure ensemble. Bars indicate

the coordinate error to the test crystal of the closest train crystal, the closest training sample, the closest AF2 model and the closest VAE generated sample.

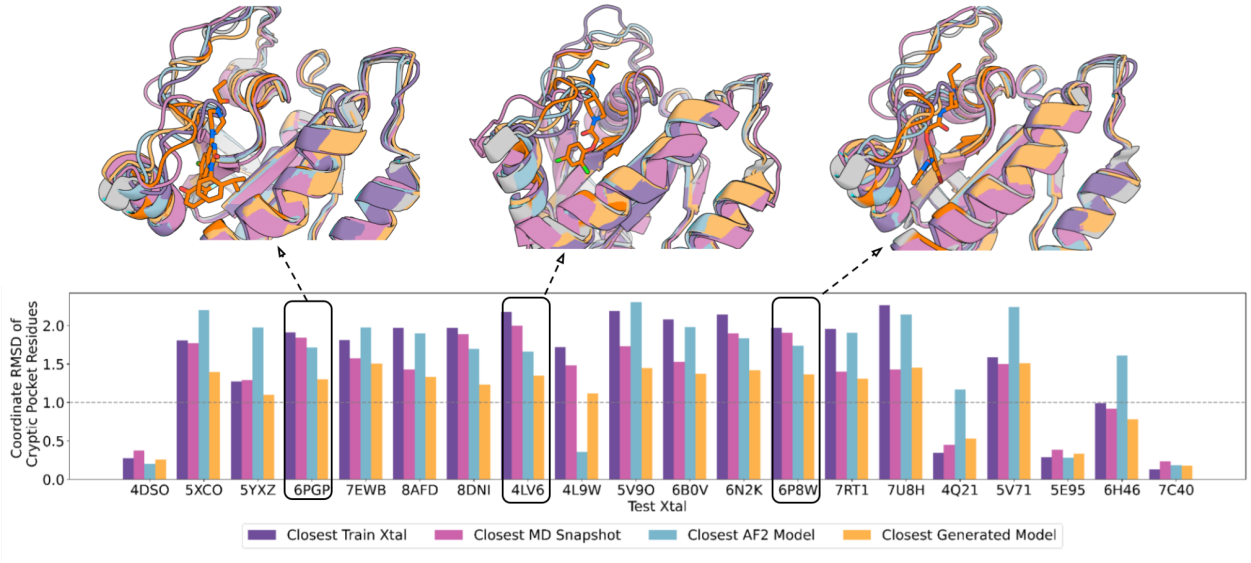

Figure 4. VAE sampling of K-Ras cryptic pocket geometry. As in Fig 3, but with the coordinate error to the test crystal structure computed over only the binding site residues (defined as the residues within 5 angstroms of the ligand binding pocket). The structural superimpositions (top) show the ligand inhibitor docked in only the target crystal, where the cryptic binding pocket and the ligand are highlighted in orange on the target crystal structure.

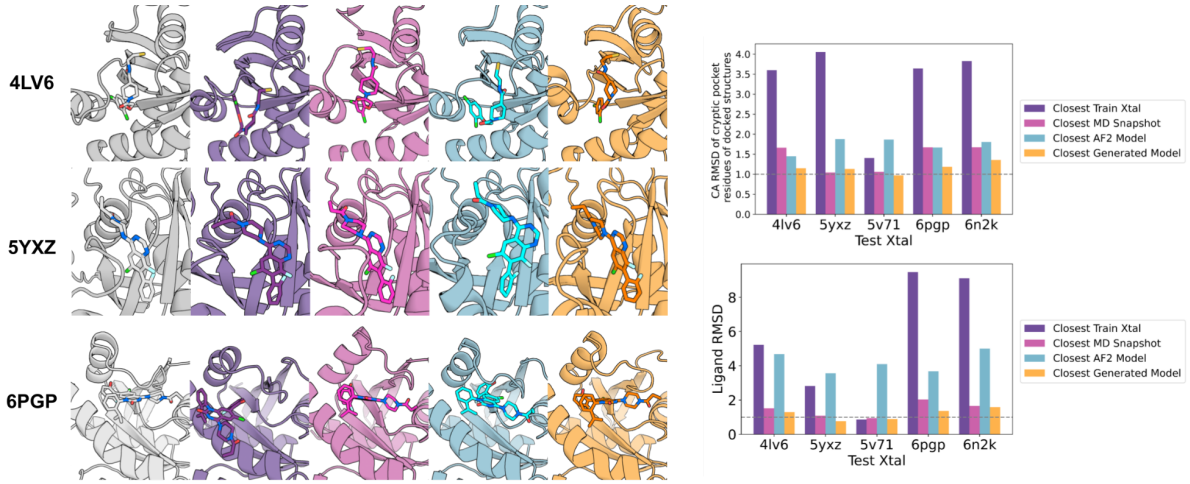

Figure 5. Small molecule docking into VAE generated ensembles. Ligands from held out crystal structures were docked into protein conformers using GA-ligand dock. Left: the held out crystal structure complex (column 1) and the closest docked complex (in terms of RMSD over the ligand) among the training set crystal structures (column 2), the MD snapshots (column 3), the AlphaFold models (column 4) , and the VAE ensembles (column 5). The closest RMSDs of C-alpha coordinate RMSD of the cryptic pocket of docked structures and lowest RMSD over ligand atoms (ligand RMSD) are indicated on the bar charts on the right.

Discussion

Our VAE based sampling approach enables extrapolation from combinations of MD simulation snapshots starting from multiple known crystal structures to generate ensembles of conformers more closely sampling held out crystal structures. The ensembles sample alternative ligand binding site geometries sufficiently accurately to enable docking of small molecule ligands. Our approach provides a way to generalize from multiple classical MD simulation trajectories from different crystal structure starting points to generate an effectively unlimited number of plausible samples with very low computational cost. We go beyond previous studies using VAEs to model the space sampled by MD simulations by taking advantage of the sophisticated understanding of protein sequence-structure relationships implicit in the AF2 and RF deep neural networks in two ways: first, we use the AF predicted distance distributions to focus the latent space sampling on regions consistent with the amino acid sequence, and second, we use RF to generate 3D coordinates from the output distance maps which ensures physical realism and local sequence-structure compatibility.

There are clear paths forward for improving our approach. First, the reconstruction error of ~1Å for the known crystal structures is reasonable, but the challenge is that the differences between many of the different conformations are also of this order, limiting the ability of our approach to really precisely sample alternative states. VAE architectures with still lower reconstruction errors would improve our method, as could fine-tuning the trained VAE on the FAPE loss (we did not observe this in preliminary tests, but this warrants further exploration). Second, while the AF2 CCE metric provides a reasonable guidepost, AF2 is trained to generate single structures, and hence the use of this measure to guide sampling could limit diversity. Better results could be obtained by minimizing towards a predicted ensemble of structures for a given target or subsampling the target MSA for RoseTTAFold structure generation (Meller et al., 2023) to introduce more diversity in output structures. Despite these limitations, our results show the utility of deep generative models for modeling the conformational ensembles that determine protein function and drugability.

Methods

1. Input data setup and incremental learning:

For the input dataset, we began by selecting distinct K-Ras conformations deposited in the PDB that are at least an angstrom away from each other as our 'training set crystal structures'. In addition to the RMSD cut-off filter, we also selected conformations that had a deposited / known inhibitor. We selected 20 K-Ras structures with these criteria. We ran MD simulations for 10 ns starting with each K-Ras crystal structure and selected every 50 ps snapshots from 5 independent trajectories, giving a total of 1000 MD snapshots for each starting structure. AMBER19SB force field (Tian et al., 2020) with TIP3P water model (Jorgensen et al., 1983) was used in a periodic boundary box. Langevin dynamics was run at a constant temperature of 300K and pressure of 1 atm. For each target crystal, the training data consisted of MD snapshots of the training set crystal structures that were at least an angstrom away from it. The final 20 K-Ras conformations that we chose were: 4DSO, 5XCO, 5YXZ, 6PGP, 7EWB, 8AFD, 8DNI, 4LV6, 4L9W, 5V9O, 6B0V, 6N2K, 6P8W, 7RT1, 7U8H, 4Q21, 5V71, 5E95, 6H46 and 7C40. All 3D structures were converted to 2D template features from RoseTTAFold (Baek et al., 2021) which consists of Cb distances and orientations. We chose to use the raw distance and orientation values for training the model for a more interpretable latent space.

After the first round of training using only MD snapshots as the training data, we then generated 3000 samples from the latent space that were optimized for the score metric and passed the diversity filter (following the protocol laid out in the sampling methods section). These 3000 generated structures were then concatenated on the initial MD snapshot training set to form an 'incremental learning' training set of structures for the model. Using this new set, for each target, the training runs were set up again from scratch. Incremental learning in this case benefits the VAE by providing a larger and more diverse set of structures for exploration, improving representation of structural diversity, refining metric optimization, and ultimately increasing the accuracy of the generated samples to the target crystal.

2. Soft Introspective VAE objective and training

We found best results using a Soft-Introspective VAE architecture (Daniel et al., 2020) which has been shown to have higher output resolution than the vanilla VAE (Kingma et al., 2014). The objective function of this model, along with the traditional VAE objective function of reconstruction loss and KL divergence, has adversarial losses incorporated, like GANs (Goodfellow et al., 2014) but is trained introspectively. In the case of SI-VAEs, the encoder is the implicit 'discriminator' where it is induced to distinguish, through the ELBO (evidence lower bound) (Kingma et al., 2014) values that it assigns to the real and generated samples. The decoder is the 'generator' where it is induced to generate samples to 'fool' the encoder (discriminator). However, unlike GANs, the SI-VAE model does not converge to the data distribution, but to an entropy-regularized version of it (Daniel et al., 2020) .

Using default parameters from Daniel *et al.,* 2020, encoder was trained with the following objective (Equation 1):

$$L_{encoder}(x, z) = s.(\beta_{rec} L_r(x) + \beta_{kl} KL(x)) + \frac{1}{2} exp(-2s.(\beta_{rec} L_r(Dec(z)) + \beta_{neg} KL(Dec(z))))$$

where $L_r(x)$ = reconstruction loss, s = 2, β_{rec} = 10, β_{kl} = 1e-3, β_{neg} = latent dimension = 256 and Dec = trained decoder of soft-introspective VAE.

The decoder was optimized using the following objective (Equation 2):

$$L_{decoder}(x,z) = s\,.\,\beta_{rec}\,L_r(x) + s\,.\,(\beta_{kl}\,KL\,(Dec(z)) + \gamma_r\,.\,\beta_{rec}\,L_r\,(Dec(z)))$$

where $L_r(x)$ = reconstruction loss, s = 2, β_{rec} = 10, β_{kl} = 1e-3 and γ_r = 1.0

The reconstruction loss was the mean-squared-error loss over all distances and orientations on the decoded template features from the model. The model was optimized using individual optimizers for the encoder and decoder, both of which were initialized with Adam (β1 = 0.9, β2 = 0.999) with learning rate 1e−3, with an effective batch size of 64. The encoder and decoder were made up of 3 ResNet blocks with 2D convolutional layers and the latent space was kept at a constant of 256 dimensions for all targets.

3. Sampling in latent space through gradient optimization of score metric (CCE)

To obtain the optimized structures using the trained decoder, we used gradient optimization in the latent space. We first randomly sample *n* numbers from the standard Gaussian distribution (mean=0, std=1) with dimension equal to that of the latent space. The initialized latent space coordinates are set to be trainable. Each sample is then decoded into its respective template features and Cb distances are discretized through radial basis function to ensure back propagation. The score metric we chose to optimize is the minimum categorical cross-entropy (CCE) among all 5 AF2 predicted Cb distograms of the target structure and the generated Cb distances. The Adam optimizer modifies the latent space sample to minimize this score metric. This process is repeated until convergence. To ensure that diversity is maintained, the latent space coordinates are restricted to explore only *d* (=10) euclidean distance in the latent space from their initial starting coordinates. The overall goal of this exploration technique is to search the latent space to find a better solution near the initial randomly generated coordinates. The final, converged latent space coordinates are decoded into their respective template features and passed into RoseTTAFold, along with the target MSA for structural modeling.

4. Docking Protocol

For each docking case, the inhibitor ligand was docked to the receptor model using the protein-ligand docking method Rosetta GALigandDock (Park et al., 2021). The ligand atomic coordinates found in complex crystal structures were extracted and used to prepare for ligand docking. The ligands were protonated and the AM1-BCC partial charges were calculated using the tools provided by openbabel, Antechamber in the AMBER suite, and UCSF Chimera (Pettersen et al., 2004). The ligand information was converted to the parameter format that is compatible with the Rosetta generic potential (*GenFF* (Park et al., 2021)). The initial position of the ligand to initiate docking was determined by superimposing the complex crystal structure to the sampled protein backbone. Protein-ligand docking was performed by allowing the side chains that are within 6A from the ligand to be flexible. The receptor models were optimized in advance using Rosetta FastRelax with high constraints on each backbone. We ran 20 parallel docking runs for each receptor model and ligand pair, and the combined results were analyzed, where the best scoring generated sample was compared to best scoring models of the training set, training crystals and AlphaFold models.

Acknowledgements and Disclosure of Funding

We would like to thank Doug Tischer, Ivan Anishchanka, Sam Pellock for helpful comments and suggestions. This work was supported by Microsoft (S.M. M.B., D.B., and generous gifts of Azure computing time), Eric and Wendy Schmidt by recommendation of the Schmidt Futures (H.P.), The Washington Research Foundation, Innovation Fellows Program (G.R.L), The Defense Threat Reduction Agency (G.R.L), The Open Philanthropy Project Improving Protein Design Fund (G.R.L), The Audacious Project at the Institute for Protein Design (D.B.), a gift from Amgen (D.B.) and the Howard Hughes Medical Institute (D.B.).

References:

1. Anand, N., & Huang, P. S. (2018). Generative modeling for protein structures. Advances in Neural Information Processing Systems.
2. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V, van Dijk, A. A., Ebrecht, A. C., … Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science, eabj8754. https://doi.org/10.1126/science.abj8754.
3. Beglov, D., Hall, D. R., Wakefield, A. E., Luo, L., Allen, K. N., Kozakov, D., Whitty, A., & Vajda, S. (2018). Exploring the structural origins of cryptic sites on proteins. Proceedings of the National Academy of Sciences of the United States of America, 115(15), E3416–E3425.
4. Cimermancic, P., Weinkam, P., Rettenmaier, T. J., Bichmann, L., Keedy, D. A., Woldeyes, R. A., Schneidman-Duhovny, D., Demerdash, O. N., Mitchell, J. C., Wells, J. A., Fraser, J. S., & Sali, A. (2016). CryptoSite: Expanding the Druggable Proteome by Characterization and Prediction of Cryptic Binding Sites. Journal of Molecular Biology, 428(4), 709–719.
5. Daniel, T., & Tamar, A. (2020). Soft-IntroVAE: Analyzing and Improving the Introspective Variational Autoencoder. http://arxiv.org/abs/2012.13253.
6. Ding,F. and Dokholyan,N.V. (2006) Emergence of protein fold families through rational design. PLoS Comput. Biol., 2, 725–733.
7. Eguchi, R. R., Anand, N., Choe, C. A., & Huang, P.-S. (2020). IG-VAE: GENERATIVE MODELING OF IMMUNOGLOBULIN PROTEINS BY DIRECT 3D COORDINATE GENERATION. BioRxiv.
8. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. http://arxiv.org/abs/1406.2661.
9. Ingraham, J., Garg, V. K., Barzilay, R., & Jaakkola, T. (2019). Generative models for graph-based protein design. Advances in Neural Information Processing Systems.
10. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics, 79(2), 926–935.
11. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature. https://doi.org/10.1038/s41586-021-03819-2.
12. Kimura, S. R., Hu, H. P., Ruvinsky, A. M., Sherman, W., & Favia, A. D. (2017). Deciphering Cryptic Binding Sites on Proteins by Mixed-Solvent Molecular Dynamics. Journal of Chemical Information and Modeling, 57(6), 1388–1401.
13. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings.
14. Larson, S.M. et al. (2002) Thoroughly sampling sequence space: large-scale protein design of structural ensembles. Prot. Sci., 11, 2804–2813.
15. Liu, D., Mao, Y., Gu, X., Zhou, Y., & Long, D. (n.d.). Unveiling the “invisible” druggable conformations of GDP-bound inactive Ras. https://doi.org/10.1073/pnas.2024725118/-/DCSupplemental.
16. Mandell,D.J. et al. (2009) Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat. Methods, 6, 551–552.
17. Meller, A., De Oliveira, S., Davtyan, A., Abramyan, T., Bowman, G. R., & van den Bedem, H. (2023). Discovery of a cryptic pocket in the AI-predicted structure of PPM1D phosphatase explains the binding site and potency of its allosteric inhibitors. Frontiers in Molecular Biosciences, 10, 1171143.
18. Meller, A., Bhakat, S., Solieva, S., & Bowman, G. R. (2023). Accelerating Cryptic Pocket Discovery Using AlphaFold. Journal of Chemical Theory and Computation. https://doi.org/10.1021/acs.jctc.2c01189

19. Park, H., Zhou, G., Baek, M., Baker, D., & Dimaio, F. (2021). Force Field Optimization Guided by Small Molecule Crystal Lattice Data Enables Consistent Sub-Angstrom Protein-Ligand Docking. Journal of Chemical Theory and Computation, 17(3). https://doi.org/10.1021/acs.jctc.0c01184.
20. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. Journal of Computational Chemistry, 25(13). https://doi.org/10.1002/jcc.20084.
21. Smith,C.A. and Kortemme.,T. (2008) Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. J. Mol. Biol., 380, 742–756. 2008;.
22. Spencer-Smith R, O'Bryan JP. Direct inhibition of RAS: Quest for the Holy Grail?. Semin Cancer Biol 2019;54:138–48. https://doi.org/10.1016/j.semcancer.2017.12.005.
23. Sun, Z., Wakefield, A. E., Kolossvary, I., Beglov, D., & Vajda, S. (2020). Structure-Based Analysis of Cryptic-Site Opening. Structure , 28(2), 223–235.e2.
24. Tian, C., Kasavajhala, K., Belfon, K. A. A., Raguette, L., Huang, H., Migues, A. N., Bickel, J., Wang, Y., Pincay, J., Wu, Q., & Simmerling, C. (2020). ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. Journal of Chemical Theory and Computation, 16(1), 528–552.
25. Tian, H., Jiang, X., Trozzi, F., Xiao, S., Larson, E. C., & Tao, P. (2021). Explore Protein Conformational Space With Variational Autoencoder. *Frontiers in Molecular Biosciences*, *8*. https://doi.org/10.3389/fmolb.2021.781635.
26. Tsuchiya, Y., Taneishi, K., & Yonezawa, Y. (2019). Autoencoder-Based Detection of Dynamic Allostery Triggered by Ligand Binding Based on Molecular Dynamics. *Journal of Chemical Information and Modeling*, *59*(9), 4043–4051. https://doi.org/10.1021/acs.jcim.9b00426.
27. Vajda, S., Beglov, D., Wakefield, A. E., Egbert, M., & Whitty, A. (2018). Cryptic binding sites on proteins: definition, detection, and druggability. Current Opinion in Chemical Biology, 44, 1–8.
28. Vijayan, R. S. K., He, P., Modi, V., Duong-Ly, K. C., Ma, H., Peterson, J. R., Dunbrack, R. L., Jr, & Levy, R. M. (2015). Conformational analysis of the DFG-out kinase motif and biochemical profiling of structurally validated type II inhibitors. Journal of Medicinal Chemistry, 58(1), 466–479.

Supplementary information:

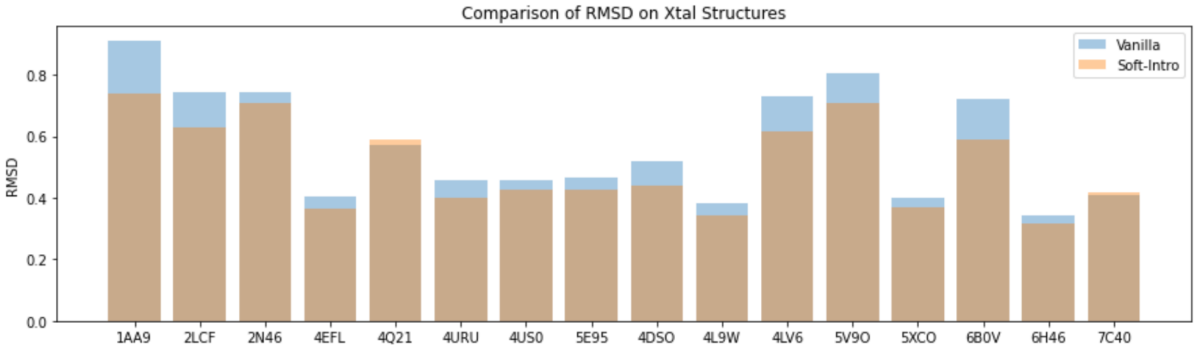

Figure S1. Comparison of reconstruction performance between vanilla VAE and soft-introspective VAE. Distance RMSD comparison of reconstruction of a different set of K-Ras training crystals from similarly trained vanilla VAE and soft-introspective VAE.

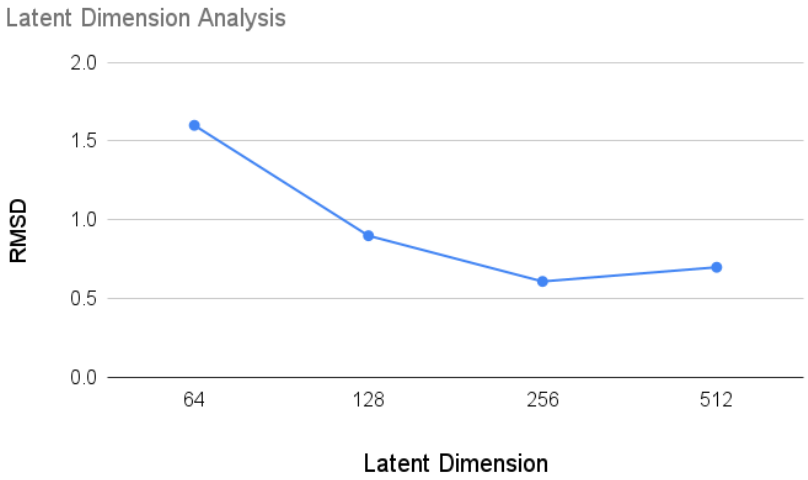

Figure S2. Graph illustrating the relationship between latent dimension and RMSD. Increasing the dimension (x-axis) initially leads to a significant decrease in mean RMSD calculated over training data (y-axis), indicating improved data representation. However, the graph reaches an elbow point (256 dimensions) where further dimension expansion yields diminishing returns, plateauing the RMSD reduction.

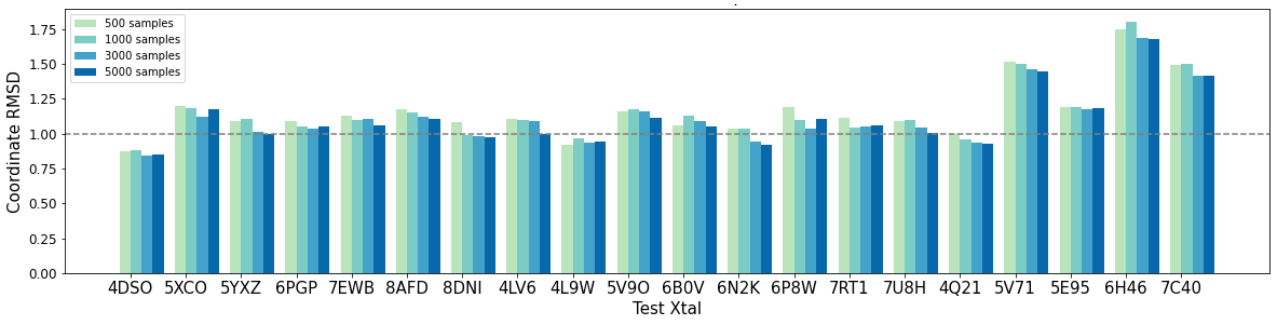

Figure S3. Relationship between increasing number of samples generated in latent space and closest coordinate RMSD to target. Each target is associated with a different number of samples generated in the latent space, and the corresponding Closest Coordinate RMSD to the target crystal is plotted. More samples result in lower RMSD until a threshold is reached, indicating improved accuracy.