AlphaSAXS: Reconstructing Protein Structure with Physiologically Relevant Conformations from Small Angle X-ray Scattering Data

Feng Yu[1*], Stephanie Prince[2*], Andrew Tritt[2], Kanupriya Pande[1,2], Greg L. Hura[1], Oliver Rübel[2], Susan E. Tsutakawa[1]
[1]Biosciences Division, Lawrence Berkeley National Laboratory
[2]Scientific Data Division, Lawrence Berkeley National Laboratory

Abstract

AlphaFold has revolutionized structural biology with accurate protein predictions, yet challenges remain due to the dynamic nature of proteins. Over 40% of human proteins have flexible regions crucial in diseases like Alzheimer's and COVID-19. To overcome AlphaFold's limitations, we integrated small-angle X-ray scattering (SAXS) data, leveraging its ability to provide structural insights on flexible macromolecules. Using computationally generated SAXS data to inform network inputs during fine-tuning and inference, we enhanced AlphaFold to integrate SAXS data to guide conformation predictions for flexible protein regions. This approach can advance our understanding of experimentally guided structural prediction and provides a potential solution for improving computational prediction of physiologically relevant conformations.

Introduction

Despite the accomplishment of AlphaFold[1], protein structure prediction remains a challenge for proteins with multiple conformations. One limitation of AlphaFold stems from its training dataset, as flexible protein regions are missing from X-ray crystallography and cryo-EM within the RCSB Protein Data Bank (PDB). To address this limitation, we adapted AlphaFold's architecture to integrate small-angle X-ray scattering (SAXS) data to enhance the prediction of conformations of flexible regions using experimental data. SAXS is a solution-based technique that measures the scattering of X-rays to provide information on the shape, size, and structural characteristics of macromolecules in solution. Similar to X-ray crystallography, the scattering intensity of SAXS can be mapped from reciprocal space to real space to obtain a pairwise distance distribution of the protein residues. Here we report a model that modifies AlphaFold architecture to incorporate SAXS pairwise distance distributions and uses this physiologically relevant conformational data to guide inference. We observed that our model results in improved accuracy and improvement in the conformational flexibility of predicted structures, proving the strategy's success.

Methods

SAXS-guided Attention Modules. To integrate SAXS data while leveraging the existing AlphaFold architecture, we introduced SAXS experimental data into the two key inputs to the network, the pair representation and multiple sequence alignment (MSA). We targeted the pair representation because SAXS data reflects pairwise distances between residues, and similarly, the pair representation encodes information about the relationship between pairs of residues. We also targeted the MSA representation because experimental SAXS data often contains information from a mixture of multiple conformations in solution, and previous work has demonstrated that modifying the MSA during inference can be used to sample alternative conformational states.[2,3] To incorporate SAXS information with these inputs, we added two multi-headed cross-attention modules, the SAXS-MSA-Attention and SAXS-Pair-Attention modules, to perform attention between the MSA and pair input embeddings and the p(r) probability distribution from the SAXS profiles (**Figure 1**). Specifically, the pair and msa representations are used to form the queries for the SAXS profile, and the attention weights are calculated to determine how each SAXS bin might influence each msa cluster and residue or residue pair, allowing these input embeddings to be updated based on relevant SAXS profile features. In the SAXS-MSA-Attention block, the number of heads was N_{heads}=8, and the dimension of the keys, queries, and values was c=32, and in the SAXS-Pair-Attention block the dimensions were N_{heads}=4, c=32. The output of these modules was added to the MSA and pair representational inputs to the Evoformer.

Training Data. We trained this network using training data with structures with multiple conformations. We fine-tuned the weights of our AlphaSAXS model starting with the publicly available pre-trained weights from AlphaFold[1] and used OpenFold[6] for the architecture implementation and training pipeline. We used two datasets to train two versions of our AlphaSAXS model. For our nuclear magnetic resonance (NMR) dataset, we obtained over 12,000 samples of NMR data from the Protein Data Bank (PDB). We removed any sequences with more than 70% similarity to the test dataset. We separated the conformations for NMR data with the BioPython library. For our synthetic normal mode analysis (NMA) dataset, we first obtained over 80,000 samples from the PDB70 (PDBs that have a similarity lower than 70%) training dataset. We fixed the flexible loops with BioPython and pdbfixer. To introduce extra flexibility and additional training data exemplifying multiple conformations and SAXS profiles for the same sequence, we then generated 6 additional conformations for each protein using the ProDY Python package. These synthetic structures were generated using an anisotropic network model and following a linear trajectory along the first normal mode with a maximum RMSD value of 3 with respect to the initial conformation. For initial training, we filtered out recordings with a number of residues > 256 since

during training the residue dimension is randomly cropped to length N_{res} = 256, and we wanted to ensure our SAXS distributions corresponded to the cropped structures without recalculating during training. We used ColabFold search with the MMSeqs2 clustering suite to generate the MSAs used in training.

SAXS profile generation. For effective training of our model on a larger training dataset, we decided to computationally generate SAXS profiles based on existing algorithms. We calculated the distance between each atom of the given PDB and then converted it into a pairwise distance distribution of electrons as provided in SAXS data. The algorithm has been tested against widely used SAXS software, including FOXS, GNOM, and RAW. The SAXS profiles for both the training dataset and the test dataset are prepared ahead of time using GNU parallel for each PDB. The SAXS profile of the predicted structure is calculated during training using an embedded version of the same algorithm.

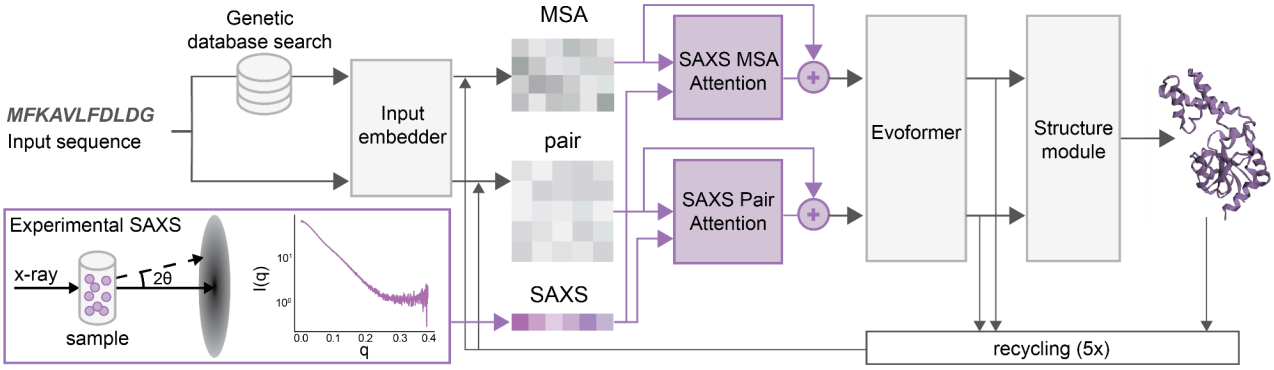

Figure 1. AlphaSAXS modifies AlphaFold2's architecture to integrate SAXS data. *SAXS information is integrated into MSA and pair embeddings via cross-attention modules prior to the Evoformer. Additions to the AlphaFold2 model are indicated in purple. We increased the number of recycling iterations since it has been shown that more recycling can improve the prediction quality at the cost of higher runtimes.*[7] *The same number of recycling iterations were used for all models, including predictions using the original AlphaFold 2 weights.*

Target Dataset. For model evaluation, we used a subset of the curated collection of apo-holo pairs of protein conformers by Saldaño et al., 2022, which provides two unique conformations for the same protein under different physiological processes. The initial collection of 91 pairs was filtered to exclude sequences with a N_{res} > 256. To focus our analysis on proteins that did not already have highly accurate predictions by the original AlphaFold, we selected protein pairs with a total pair RMSD > 4Å and an individual conformer RMSD > 2.5Å, resulting in an evaluation dataset of 40 proteins. We calculated the average Cα-RMSD change of the conformation unfavoured by AlphaFold for all protein sequences in the target dataset. Compared to the AlphaFold and MSA-subsampling strategy (implemented with the AlphaFlow[5] network), our model generally shows an improvement in average RMSD and lower standard deviation (**Figure A1, Appendix**).

Results

Alphafold fails to predict the different conformations of the same protein during binding with ligands, different solution conditions, or inherent flexibility. Here we test over pairs of protein conformations that have the same sequence to evaluate how our SAXS-guided model can predict different protein conformations with a given simulated experimental data. This demonstrates real-world use cases with our model under different biological problems.

AlphaSAXS-NMR improves prediction accuracy while matching the SAXS input data. One biological example we observed in our test dataset is β-Phosphoglucomutase (β-PGM PDB: 1ZOL/1O03). β-PGM functions rely on the conformational change of the protein, but AlphaFold can only predict one conformation. With our model, we predict the correct conformation of β-PGM with a 1.28Å RMSD that reached the original experimental accuracy of 1.9Å (**Figure 3**). We also observed that our prediction shows a significant change in the SAXS P(r) that matched the reference structure.

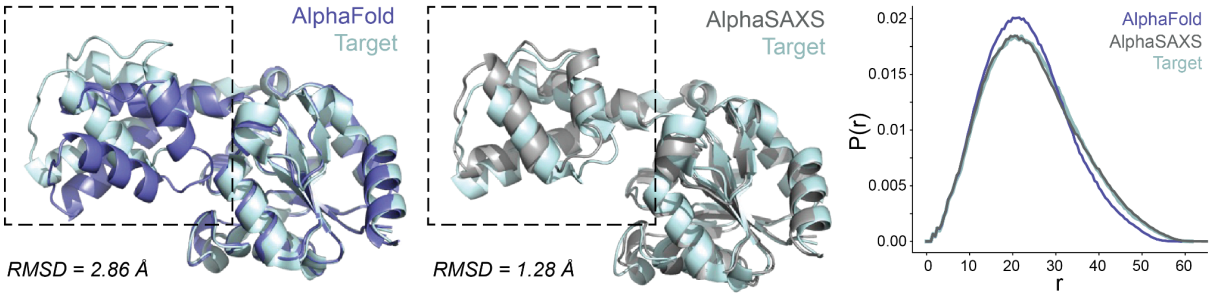

Figure 2. AlphaSAXS improves structural and SAXS prediction accuracy. *Left: AlphaFold prediction vs. target structure. Middle: AlphaSAXS prediction vs. target structure. Right: Theoretical SAXS p(r) distributions calculated from the target.*

AlphaSAXS trained with synthetically generated data increases the conformational diversity. To further improve the conformational diversity of our model's predictions, we trained the model with a synthetic dataset of conformations generated using normal mode analysis (NMA) to introduce information about potential movements of flexible protein regions. Using this approach, we observed larger conformational changes, e.g., the calcium-binding protein from *Entamoeba histolytica* (EhCaBP, PDB: 1JFJ/1JFK). Similar to β-PGM, this protein relies on conformational diversity for its signal transduction mechanism. However, AlphaFold cannot predict two distinct conformations of the protein and instead generates overly compact models. Our NMR model can reproduce a conformation similar to that of 1JFK but fails to capture the more flexible conformation observed in 1JFJ. In contrast, the NMA model produces a more extended conformation with enhanced flexibility but with some incorrect secondary structures. To determine how our AlphaSAXS model and framework impact the overall diversity of conformational predictions, we compared the variation in prediction outputs in the test set, providing the same sequence and distinct SAXS data for each conformation in an apo-holo pair. We found that both AlphaSAXS models demonstrated larger differences in the predicted structures compared to AlphaFold2, 3.6x larger median RMSD in the NMA model and 3.1x larger median RMSD in the NMR model (**Figure 3**, **Figure A1**).

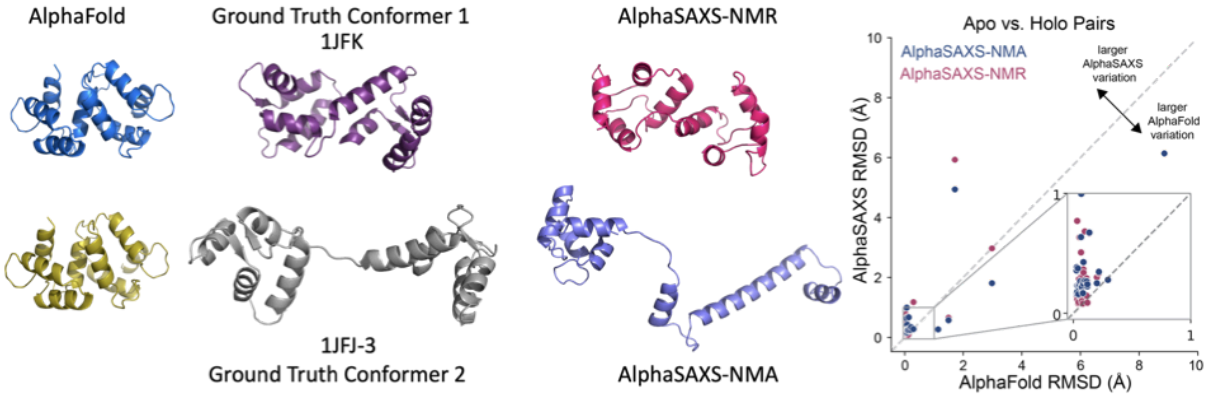

Figure 3. The NMA training dataset enables AlphaSAXS to predict highly flexible open protein conformations. *Left: AlphaFold predicts highly compact incorrect structures for both conformations. Middle Left: The two ground truth structures. Middle Right: AlphaSAXS generated more flexible structures. Right: RMSD shows the diversity of each pair of proteins with the same sequence. The left axis shows the RMSD values for the predicted outputs from the AlphaSAXS models vs. the target structures and the bottom axis shows the RMSD values for the predicted outputs from AlphaFold vs. the target structures. Compared to the AlphaFold model, AlphaSAXS demonstrates an improvement of conformational diversity while increasing the diversity.*

Discussion

In this year's critical assessment of structure prediction (CASP16), we and our collaborators organized an ensemble prediction competition to evaluate the performance of ML models for flexible protein regions with experimental data. None of the ML models generated a close prediction that matched the experimental data. Our AlphaSAXS approach is our initial effort to combine the biophysical experimental knowledge to drive the prediction of flexible proteins and demonstrates promising new directions to explore structural prediction with SAXS data. Despite this initial success in the improvement of structural RMSD and conformational diversity, there remain several areas we wish to improve upon. First, the conformational diversity is still relatively small, and future work will explore approaches to increase the variation in the prediction. Second, our work thus far has demonstrated how theoretical SAXS data can improve AlphaFold predictions, and our end goal is to apply AlphaSAXS to experimental SAXS data, generating multiple conformational predictions that reflect the underlying data distribution. Overall, our model suggests a novel strategy to integrate AlphaFold constraints with low-resolution experimental results to generate physiologically relevant protein structures.

Appendix

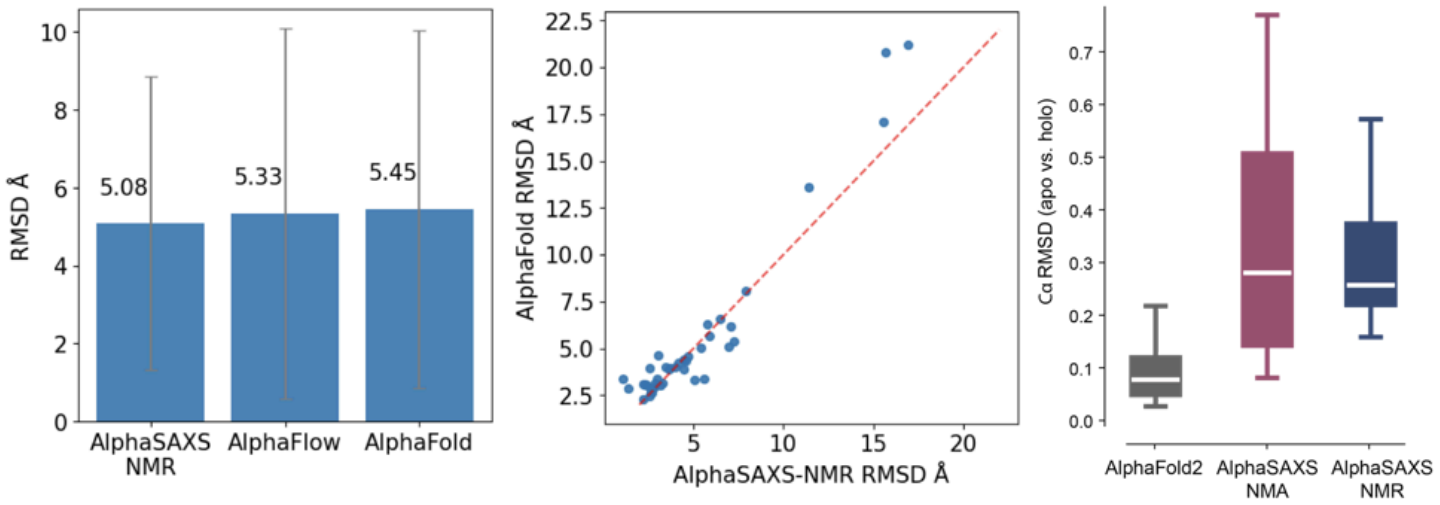

Figure A1. AlphaSAXS improves the average RMSD of the test dataset. *Left: Average barplot shows the comparison of RMSD of prediction vs. target for different models. This represents the improvement of the model accuracy. Middle: Compare RMSD change between AlphaFold and AlphaSAXS for 40 data points. Right: Pairwise RMSD between protein pairs that have the same sequence. This RMSD shows the diversity of predictions across different models.*

Table A1. Training Strategy

	Normal Mode Analysis with PDB70	NMR dataset with PDB
Hardware per GPU nodes	**1x AMD EPYC 7763 4x NVIDIA A100 (40GB)**	**1x AMD EPYC 7763 4x NVIDIA A100 (40GB)**
Nodes used for training	**64**	**64**
Size of Training Dataset	**298,025**	**214,973**
Effective Epochs	**37**	**13**
Total GPU Hours	**7,680**	**3072**

References

1. Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).
2. Diego del Alamo, Davide Sala, Hassane S Mchaourab, Jens Meiler. Sampling alternative conformational states of transporters and receptors with AlphaFold2 eLife 11:e75751 (2022).
3. Wayment-Steele, H.K., Ojoawo, A., Otten, R. et al. Predicting multiple conformations via sequence clustering and AlphaFold2. Nature 625, 832–839 (2024).
4. Saldaño T, Escobedo N, Marchetti J, Zea DJ, Mac Donagh J, Velez Rueda AJ, Gonik E, García Melani A, Novomisky Nechcoff J, Salas MN, Peters T, Demitroff N, Fernandez Alberti S, Palopoli N, Fornasari MS, Parisi G. Impact of protein conformational diversity on AlphaFold predictions. Bioinformatics. 2022 May 13;38(10):2742-2748.
5. Jing, B., Berger, B., & Jaakkola, T. . AlphaFold meets flow matching for generating protein ensembles. arXiv preprint arXiv:2402.04845 (2024).
6. Ahdritz, G., Bouatta, N., Floristean, C. et al. OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. Nat Methods 21, 1514-1524 (2024).
7. Mirdita, M., Schütze, K., Moriwaki, Y. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* 19, 679–682 (2022). https://doi.org/10.1038/s41592-022-01488-1

Acknowledgments

This research used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy Office of Science User Facility using NERSC awards ERCAP0030518 and ERCAP0034111.