Background

Molecular dynamics (MD) simulations are critical for understanding proteins *in action*.

- **All-atom molecular dynamics** protein simulations assisted in understanding disease and drug design for COVID-19
  - Discovered property of spike glycans to modulate infectivity (Casalino et al., 2020)
  - Engineered drug ML2006a4, high affinity for main protease (Westberg et al., 2024)

Figure created by student author.

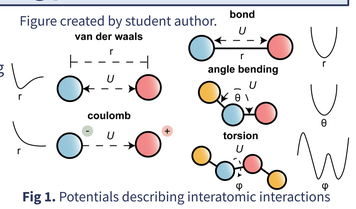


Fig 1. Potentials describing interatomic interactions allow simulations to predict molecular movement.

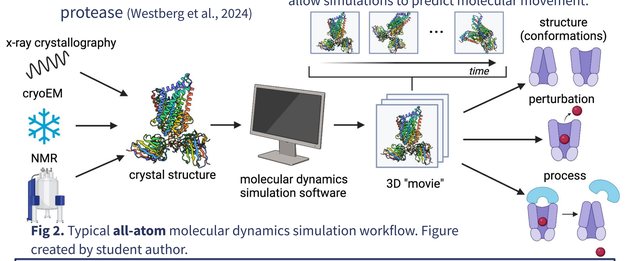


Fig 2. Typical **all-atom** molecular dynamics simulation workflow. Figure created by student author.

Extensive all-atom simulations remain largely *inaccessible*.

- AA simulations of larger proteins require extensive computing power and are time consuming
  - Calculates ~10^6-~10^9 interactions for large proteins with thousands of atoms, at each time step (1 fs), so since biomolecular processes take place on ns, µs, or ms → billions of time steps
- COVID-19 spike protein (~45,000 atoms) simulated for 1.8 ms is estimated to take **~13.7 years** on a consumer-accessible GPU (NVIDIA GeForce GTX 1080Ti) (Thomson et al., 2021)

Coarse-graining and backmapping speed up simulations by simplifying them.

- **Coarse-graining (CG)** into structures with fewer particles enables faster runtimes with less computing power (Kmiecik et al., 2016)
  - Constituent atom *identities* retained but *positional* information lost
- Positional information at the *atomic level* is crucial for some simulation applications (e.g. mutation effects, local conformational changes)
- **Backmapping** restores atomic detail, predicting positions of constituent atoms given bead coordinates at each time step (Wassenaar et al., 2014)

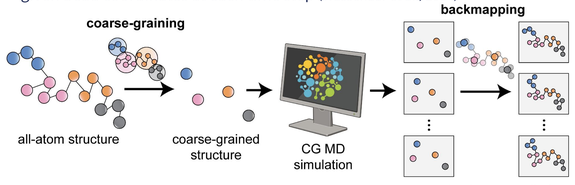


Fig 3. Typical **coarse-grained** molecular dynamics simulation workflow. Figure created by student author.

Current backmapping methods experience a *tradeoff* between efficiency and performance.

- Fewer beads yield faster, cheaper simulations, but simulations are currently restricted to larger number of beads due to limitations in backmapping (Wang et al., 2022).

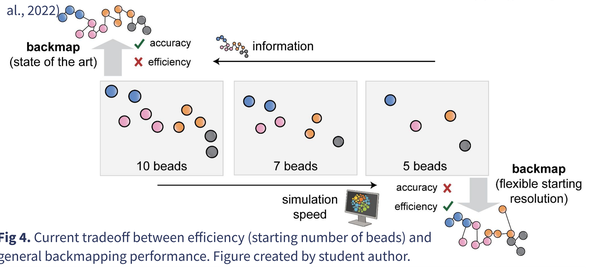


Fig 4. Current tradeoff between efficiency (starting number of beads) and general backmapping performance. Figure created by student author.

A "divide and conquer" strategy shows promise in application to similar problems.

- Inspired by developments in image super resolution, this study proposes an **iterative reconstruction** strategy

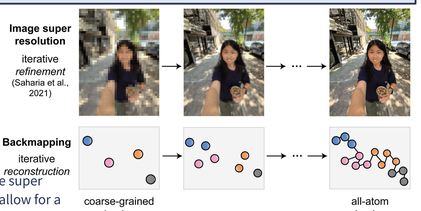


Fig 5. Parallels between image super resolution and backmapping allow for a similar approach to both problems. Figure created by student author.

Objective—apply a similar, iterative *reconstruction* framework to backmapping to **improve performance** from low resolution starting coarse-grained structures, allowing for faster simulations requiring fewer resources.

StAIR

Stepwise All-atom Iterative Reconstruction of Large Coarse-Grained Protein Structures

Erin Wong, Great Neck South High School, Great Neck, NY, 11020

Methods

StAIR Framework

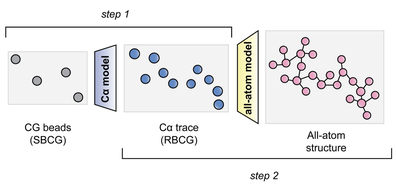
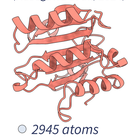


Fig 6. Schematic of StAIR framework. Two models, rather than one, are applied sequentially, backmapping first to an "*intermediate*" resolution with more beads than the initial structure. Figure created by student author.

Data

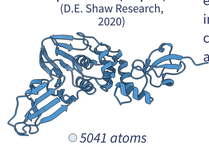
Proteins

Eukaryotic initiation factor 4E (eIF4E) (Hasegawa et al., 2024)



2945 atoms

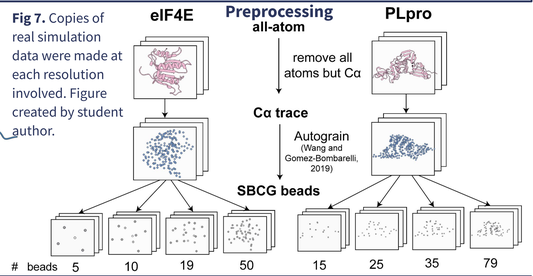
SARS-CoV-2 papain-like protease (PLpro) (D.E. Shaw Research, 2020)



5041 atoms

- 3,000 timesteps each (frames) of real all-atom simulation data for both proteins were used. Figure created by student author.

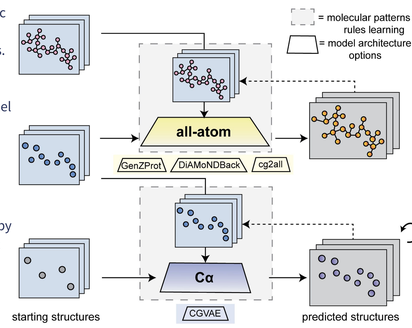
Fig 7. Copies of real simulation data were made at each resolution involved. Figure created by student author.



Training

Fig 8. Schematic illustrating training process. Component models were trained in parallel on their respective portions of the split simulation data.

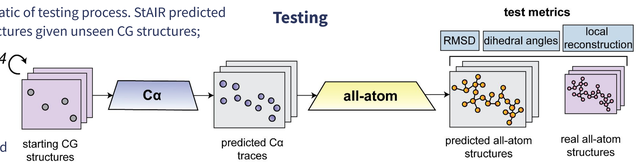
Figure created by student author.



Testing and Application

Fig 9. Schematic of testing process. StAIR predicted all atom structures given unseen CG structures; these were compared to the true simulation structures.

Figure created by student author.



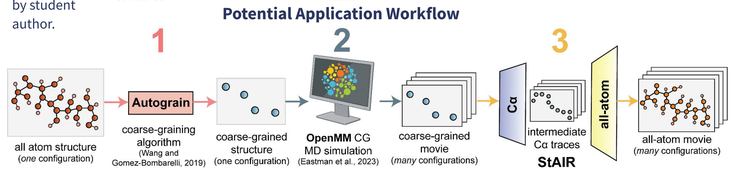


Fig 10. StAIR's role in application to a coarse-grained molecular dynamics setting. Figure created by student author.

Results and Discussion

StAIR accelerates coarse-grained simulations by up to 176x while improving backmapping quality

StAIR outperformed the current baseline across all metrics.

Figure created by student author.

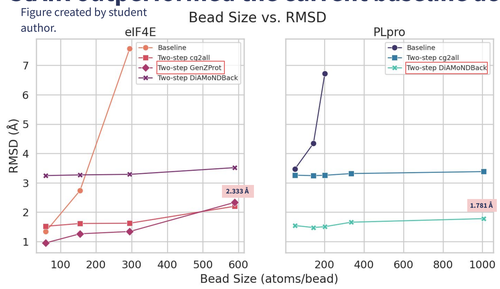


Fig 11. Scatterplots comparing StAIR and the baseline CGVAE's performance across all starting CG resolutions (bead size) quantified by root mean square deviation (RMSD) for (a) eIF4E and (b) PLpro. The performances of multiple "variants" of StAIR with different all-atom model architectures are plotted as separate trendlines. Best-performing variants for each protein are boxed in red.

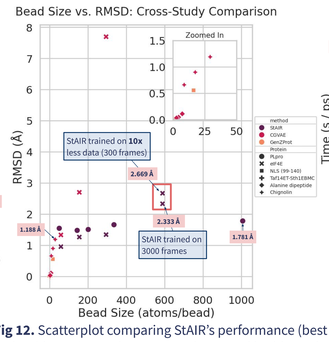


Fig 12. Scatterplot comparing StAIR's performance (best variant) in RMSD on the two proteins tested in this study to previous studies using other models for different proteins. Figure created by student author.

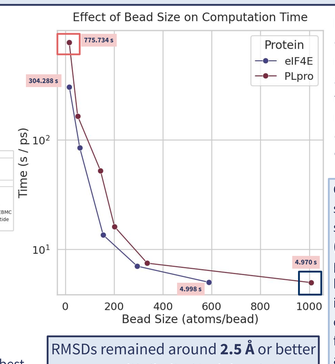


Fig 13. Scatterplot of computing time needed for 1 ps of simulation at each of the starting resolutions for both eIF4E and PLpro. Figure created by student author.

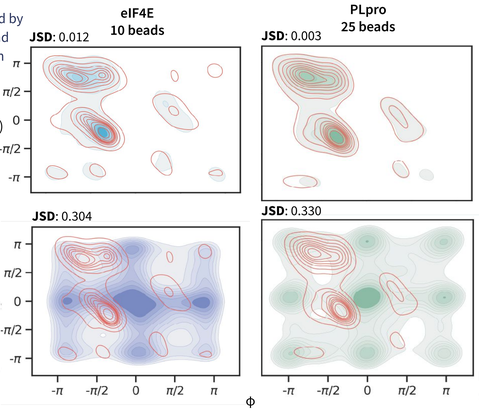
Current methods limit simulation to a bead size of **~15** atoms/bead (boxed in red) due to poor performance given larger sizes, but StAIR's improved performance enables backmapping from up to **~1000** atoms/bead (boxed in blue) (Angioletti et al., 2024; Heo and Feig, 2024; Yang and Gomez-Bombarelli, 2023).

RMSDs remained around **2.5 Å** or better even with *limited training data*, unlocking the potential to simulate faster at resolutions **up to 64x** lower.

StAIR produced more realistic protein geometries.

An iterative approach is necessary to learn the distribution of common and rarer secondary structures.

Fig 14. Ramachandran plots of dihedral angle distributions for reconstructed structures produced by StAIR (first row) and CGVAE (second row). The red outlined distribution is from reference simulation structures. Above each plot is the Jensen-Shannon Divergence (JSD) quantifying the similarity between the reconstructed and reference distributions; lower values indicate greater similarity.

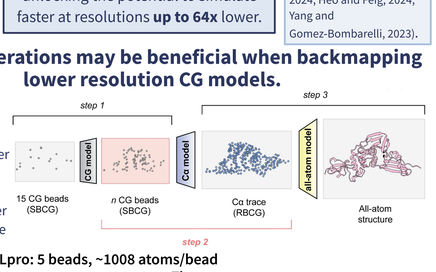


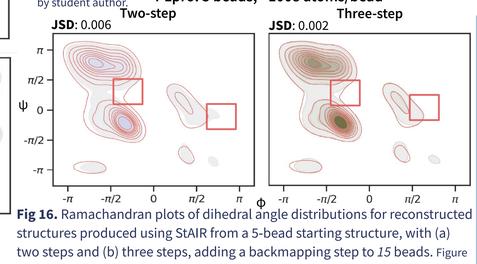
The baseline model lacked the ability to restore meaningful secondary structures, showing minimal overlap while StAIR with two steps sufficiently did so.

Figure created by student author.

Additional iterations may be beneficial when backmapping lower resolution CG models.

Fig 15. Schematic of StAIR with another step inserted prior to backmapping to Cα traces. *n* is any number of beads between the number of starting beads and the number of carbon atoms in the Cα trace. Figure created by student author.





The third backmapping step provided another constraint during backmapping and an *additional source of data* to learn patterns from (at the new SBCG resolution), thus **reducing oversampling**.

Fig 16. Ramachandran plots of dihedral angle distributions for reconstructed structures produced using StAIR from a 5-bead starting structure, with (a) two steps and (b) three steps, adding a backmapping step to *15* beads. Figure created by student author.

Results and Discussion

StAIR preserves the local structure of residues critical for protein function.

Tryptophan residues integral for eIF4E's cap binding ability were reconstructed.

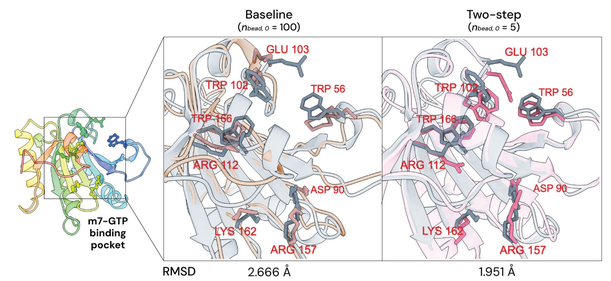


Fig 17. Zoomed-in visualizations of reconstruction quality at the eIF4E m7-GTP binding pocket. All-atom reconstruction and reference structures in the frame where two-step backmapping achieved the *lowest* RMSD are shown. Reference structures are gray while reconstructed structures are colored. Figure created by student author.

Increased model complexity resulted in improved reconstruction of PLpro's active site and highly variable BL2 ligand-binding loop.

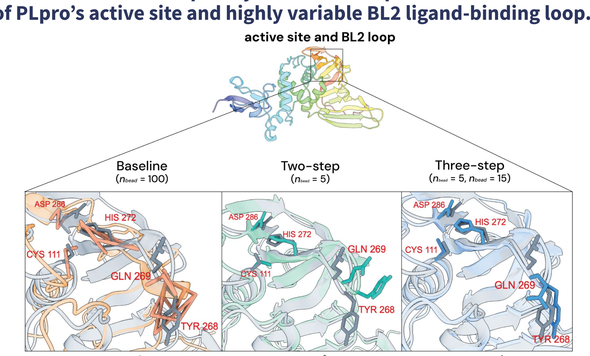


Fig 18. Zoomed-in visualizations of reconstruction quality at the selected PLpro sites of interest. Two sites are shown: the **catalytic core** and the **BL2 loop** which regulates ligand-binding and has been characterized to be highly variable in conformation (Ferreira et al., 2022). All-atom reconstruction and reference structures in the frame where three-step backmapping had the *lowest* RMSD are shown. Reference structures are gray while reconstructed structures are colored. Figure created by student author.

Limitations

- Lacks **chemical transferability**; the component models of StAIR must be newly trained for each protein first before application
- Tested and applied to two *globular* proteins → might not test full extent of StAIR to be applied to diverse structures (i.e. knotted and intrinsically disordered proteins even larger in size)

Conclusions

1. Improved reconstruction quality from all starting resolutions, even with *limited* data → **extensive characterization of novel protein structures**
2. 176x speed up in simulations (weeks to hours) → **large protein simulations on standard hardware**
3. Preservation of local geometry → **drug design and understanding protein involvement in disease**

Future Directions

- Testing larger proteins (500-1000 residues) by extending StAIR to use additional iterations
- Incorporate alternative coarse-graining methods to RBCG (Cα traces) for constructing the intermediate resolution structures like Martini and PRIMO (Monticelli et al., 2008; Kar et al., 2013)
- Use a diffusion model architecture to natively breakdown iterative reconstructions into each step, within a single model

References

Angioletti, D., Raniolo, S., & Limongelli, V. (2024). HEroBM: A deep equivariant graph neural network for universal backmapping from coarse-grained to all-atom representations (No. arXiv:2404.16911). arXiv. http://arxiv.org/abs/2404.16911

Casalino, L., Gaieb, Z., Goldsmith, J. A., Hjorth, C. K., Dommer, A. C., Harbison, A. M., Fogarty, C. A., Barros, E. P., Taylor, B. C., McLellan, J. S., Fadda, E., & Amaro, R. E. (2020). Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. ACS Central Science, 6(10), 1722–1734. https://doi.org/10.1021/acscentsci.0c01056

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., & Pande, V. S. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLOS Computational Biology, 13(7), e1005659. https://doi.org/10.1371/journal.pcbi.1005659

D. E. Shaw Research. (2020). Molecular Dynamics Simulations Related to SARS-CoV-2. D. E. Shaw Research Technical Data. https://www.deshawresearch.com/downloads/download_trajectory_sarscov2.cgi/

Ferreira, G., Pillaiyar, T., Hirata, M., Pess, A., & Kronenberger, T. (2022). Inhibitor induced conformational changes in SARS-COV-2 papain-like protease. Scientific Reports, 12. https://doi.org/10.1038/s41598-022-15183-y

Hasegawa, K., Burzachiello, J., Connolly, L., Kementzidis, G., Niu, Z., Papadopoulos, E., Aktas, B. H., & Deng, Y. (2024). Structural and Dynamical Analyses of Apo and Cap-binding eIF4E: An in silico Study (p. 2024.05.18.594816). bioRxiv. https://doi.org/10.1101/2024.05.18.594816

Heo, L., & Feig, M. (2024). One bead per residue can describe all-atom protein structures. Structure (London, England: 1993), 32(1), 97–111.e6. https://doi.org/10.1016/j.str.2023.10.013

Hollingsworth, S. A., & Dror, R. O. (2018). Molecular dynamics simulation for all. Neuron, 99(6), 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011

Jones, M. S., Shmilovich, K., & Ferguson, A. L. (2023). DiAMoNDBack: Diffusion-Denoising Autoregressive Model for Non-Deterministic Backmapping of Cα Protein Traces. Journal of Chemical Theory and Computation, 19(21), 7908–7923. https://doi.org/10.1021/acs.jctc.3c00840

Kar, P., Gopal, S. M., Cheng, Y.-M., Predeus, A., & Feig, M. (2013). PRIMO: A Transferable Coarse-Grained Force Field for Proteins. Journal of Chemical Theory and Computation, 9(8), 3769–3788. https://doi.org/10.1021/ct400230y

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., & Kolinski, A. (2016). Coarse-Grained Protein Models and Their Applications. Chemical Reviews, 116(14), 7898–7936. https://doi.org/10.1021/acs.chemrev.6b00163

Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., & Marrink, S.-J. (2008). The MARTINI Coarse-Grained Force Field: Extension to Proteins. Journal of Chemical Theory and Computation, 4(5), 819–834. https://doi.org/10.1021/ct7002026

Thomson, E. C., Rosen, L. E., Shepherd, J. G., Spreafico, R., da Silva Filipe, A., Wojcechowsky, J. A., Davis, C., Piccoli, L., Pascall, D. J., Dillen, J., Lytras, S., Czudnochowski, N., Shah, R., Meury, M., Jesudason, N., De Marco, A., Li, K., Bassi, J., O'Toole, A., … Snell, G. (2021). Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. Cell, 184(5), 1171-1187.e20. https://doi.org/10.1016/j.cell.2021.01.037

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., & Norouzi, M. (2023). Image Super-Resolution via Iterative Refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4), 4713–4726. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2022.3204461

Wang, W., & Gómez-Bombarelli, R. (2019). Coarse-graining auto-encoders for molecular dynamics. Npj Computational Materials, 5(1), 125. https://doi.org/10.1038/s41524-019-0261-5

Wang, W., Xu, M., Cai, C., Miller, B. K., Smidt, T., Wang, Y., Tang, J., & Gomez-Bombarelli, R. (2022). Generative Coarse-Graining of Molecular Conformations. Proceedings of the 39th International Conference on Machine Learning, 23213–23236. https://proceedings.mlr.press/v162/wang22g.html

Wassenaar, T. A., Pluhackova, K., Böckmann, R. A., Marrink, S. J., & Tieleman, D. P. (2014). Going Backward: A Flexible Geometric Approach to Reverse Transformation from Coarse Grained to Atomistic Models. Journal of Chemical Theory and Computation, 10(2), 676–690. https://doi.org/10.1021/ct400617g

Westberg, M., Su, Y., Zou, X., Huang, P., Rustagi, A., Garhyan, J., Patel, P. B., Fernandez, D., Wu, Y., Hao, C., Lo, C.-W., Karim, M., Ning, L., Beck, A., Saenkham-Huntsinger, P., Tat, V., Drelich, A., Peng, B.-H., … Einav, S. … Lin, M. Z. (2024). An orally bioavailable SARS-CoV-2 main protease inhibitor exhibits improved affinity and reduced sensitivity to mutations. Science Translational Medicine, 16(738), eadi0979. https://doi.org/10.1126/scitranslmed.adi0979

Yang, S., & Gomez-Bombarelli, R. (2023). Chemically Transferable Generative Backmapping of Coarse-Grained Proteins. Proceedings of the 40th International Conference on Machine Learning, 39277–39298. https://proceedings.mlr.press/v202/yang23e.html

extra figures

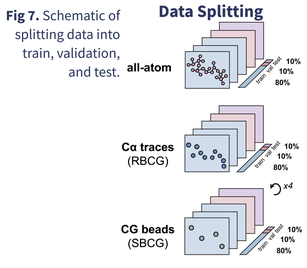


Fig 7. Schematic of splitting data into train, validation, and test.