Challenges and Guidelines in Deep Generative Protein Design: Four Case Studies

Tianyuan Zheng [1] **Alessandro Rondina** [2] **Gos Micklem** [3] **Pietro Liò** [4]

Abstract

Deep generative models show promise for *de novo* protein design, yet reliably producing designs that are geometrically plausible, evolutionarily consistent, functionally relevant, and dynamically stable remains challenging. We present a deep generative modeling pipeline for early *de novo* design of monomeric proteins, based on Score Matching and Flow Matching. We apply this pipeline to four diverse protein families with an adaptable evaluation protocol. Generated structures display realistic, clash-free conformations enriched with family-specific features, while the designed sequences preserve essential functional residues while retaining variability. Molecular dynamics and binding simulations show dynamic stability, with wild-type-like binding pockets that interact favorably with family-specific ligands. These results provide practical guidelines for integrating generative models into protein design workflows.

1. Introduction

Protein functions and specificities are dictated by their complex structures. Over the past 60 years, we have progressed from viewing protein design as unattainable to achieving complete artificial design and synthesis (Korendovych & DeGrado, 2020; Huang et al., 2016), with expanding applications across industries (Arunachalam et al., 2021; Kingwell, 2024; Barclay & Acharya, 2023; Victorino da Silva Amatto et al., 2021; Ali et al., 2020). However, achieving atomic-level precision remains challenging due to the nonlinear complexity of folding and the sensitivity of the function to slight changes, still demanding significant resources.

With the rapid growth of protein structure databases (Jamasb et al., 2024; Berman, 2000), various *in silico de novo* protein design methods, particularly deep generative models (Watson et al., 2023; Ingraham et al., 2023; Wu et al., 2022), have emerged. However, many methods generate structures that appear novel, diverse, and designable, yet few assess their functionality or evolutionary relevance, leaving the observations potentially as mere (Ji et al., 2023). Unclear biological functionality, unknown stability, and the lack of validation connecting these structures to known functions limit the broader application and advancement of these methods.

We tackle these challenges using a deep generative pipeline, based on diffusion-based Score Matching (SM) and Flow Matching (FM), for early *de novo* design of monomeric proteins. Guided by an adaptable evaluation protocol, we apply it to four protein families chosen for their diverse structural folds, functional roles, and rich annotations (Appendix C). Generated structures are realistic and clash-free, and structural phylogenetic analyses show that they capture family-specific features consistent with common ancestry and function. Designed sequences conserve key functional residues while allowing variability elsewhere, yielding low-similarity variants with similar functions. Molecular Dynamics (MD) simulations demonstrate dynamic stability of the designs, while forming wild-type (WT)-like binding pockets that favorably interact with family-specific ligands in docking studies. Together, these results offer practical guidelines for integrating generative models into protein design workflows.

2. Pipeline and Evaluation Protocol

2.1. Protein Backbone Generation

In Appendix B, we review the key concepts behind the approaches of Yim et al. (2023) and Bose et al. (2023) for backbone generation, using SM and FM on $\mathrm{SE}(3)$.

2.2. Geometric Plausibility

Due to steric hindrance and spatial repulsion, not all backbone dihedral angles in proteins are physically feasible or energetically favorable. We analyzed the ϕ and ψ distribution in the generated structures using Ramachandran plots (Ramachandran et al., 1963) (Figure 6).

[*]Equal contribution [1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK [2]Department of Molecular and Translational Medicine, University of Brescia, Brescia, Italy [3]Department of Genetics, University of Cambridge, Cambridge, UK [4]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK. Correspondence to: Tianyuan Zheng <tz365@cam.ac.uk>.

Proceedings of the ICML 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Vancouver, Canada. 2025.

2.3. Conserved Residue Consistency

In protein families or across species, certain residues are highly conserved, typically to ensure function but also structural stability and to support proper folding.

Determining the optimal amino acid sequence. Following Yim et al. (2023), we used ProteinMPNN (Dauparas et al., 2022) to predict ten sequences for each generated backbone. We then modeled these sequences with EMSFold (Rives et al., 2019) and compared them to the generated backbone using the TM_{score} (Zhang, 2005). The sequence with the highest TM_{score} was chosen as the optimal match.

Identifying conserved residues. Experimentally derived sequences were aligned with optimal sequences of generated backbones using Clustal Omega (Sievers et al., 2011). Given these alignments and generated structures, ConSurf (Yariv et al., 2023) reconstructed phylogenetic trees and applied Rate4Site (Pupko et al., 2002) to estimate per-residue evolutionary rates via an empirical Bayesian method (Ashkenazy et al., 2016; Mayrose, 2004).

2.4. Structural Phylogenetics

Certain applications require designing structures with low sequence similarity while retaining similar functions. However, when sequence identity falls below 30%, homology detection and evolutionary inference become challenging (Puente-Lelievre et al., 2023). Structural comparisons, which are more conserved, offer a more effective alternative (Illergård et al., 2009; Flores et al., 1993; Moi et al., 2023).

Using Q_{score} in structural phylogenetics. Malik et al. (2020) proposed the Q_{score} (Krissinel & Henrick, 2004) for structural phylogenetics, which accommodates indels and combines alignment quality with length. It compares the positions of all C_α atoms across N_{align} comparable residues in pairwise comparisons. We construct structural phylogenetic trees using $1 - Q_{\text{score}}$ as a distance measure, where higher values indicate greater structural similarity.

Using 3Di alphabet in structural phylogenetics. van Kempen et al. (2023) developed Foldseek, which encodes protein tertiary interactions using a 20-state 3D interaction (3Di) alphabet to simplify structural alignments. This approach reduces false positives and increases information density by effectively encoding conserved core regions. Leveraging Foldseek's divergence metrics, Moi et al. (2023) constructed structural phylogenetic trees based on rigid body alignment, local alignment without superposition, and sequence alignment with structural alphabets, showing that these trees outperform traditional sequence-based trees across varied evolutionary timescales (Moi et al., 2023).

2.5. Molecular Dynamics

Structures that appear reasonable may, in fact, be unstable due to molecular dynamics, water interactions, and entropy, which are not accounted for during generation. To assess the dynamic stability of these generated structures, we conducted MD simulations under physiological conditions and analyzed their time-dependent behavior.

Homology modeling of side-chains. We used homology modeling to add side-chains to the protein backbone. After determining the optimal sequence (Section 2.3), we selected template proteins with at least 45% sequence identity and a TM_{score} of 0.75 or higher using FoldSeek (van Kempen et al., 2023). Sequences were aligned with Clustal Omega, and MODELLER (Šali & Blundell, 1993) generated possible side-chain conformations using statistical potentials and rotamer libraries, with backbone fixed. We chose the final side-chain arrangement based on lowest energy and minimal steric clashes, verifying model quality with PROCHECK (Laskowski et al., 1993; 1996) and WHAT_CHECK (Hooft et al., 1996), discarding low-quality models. MODELLER's modeling leverages (a) the input alignment to position side-chains, (b) template structures to set spatial restraints that mimic contacts, (c) knowledge-based rotamer distributions to favor typical angles, and (d) an optimization process.

Simulation. Hydrogen atoms were added using Reduce2 from the computational crystallography toolbox (Grosse-Kunstleve et al., 2002), and MD simulations were performed with GROMACS (Abraham et al., 2015). Proteins were placed in an octahedral box with at least 15 Å from the edges, and intermolecular interactions were modeled using the CHARMM36 force field (July 2022) (Vanommeslaeghe et al., 2009; Vanommeslaeghe & MacKerell, 2012; Yu et al., 2012). After vacuum energy minimization, the system was solvated, neutralized to 150 mM Na^+ and Cl^-, and minimized again. Then it was heated to 310 K under NVT[1], equilibrated at 1 atm under NPT[2], and subjected to a 10 ns production run. Details are provided in Appendix I.

2.6. Protein-ligand Docking

AutoDock Vina (Trott & Olson, 2009; Eberhardt et al., 2021) was used to predict optimal binding modes between generated structures and their family-specific ligands. We performed blind docking, scanning the entire protein surface for potential ligand binding sites without prior knowledge of the pockets. The grid box covered the entire protein, and identical configurations were applied to both generated and experimentally derived structures to evaluate whether deviations in the generated samples fell within an acceptable

[1]Constant number of particles, volume, and temperature.
[2]constant number of particles, pressure, and temperature

range, thus assessing their functional viability. Details on simulation settings are provided in Appendix I.3.

3. Case Studies

3.1. Data

We assembled a dataset of monomeric proteins covering four families (β-lactamases, cytochrome c, GFP, and Ras), with varied fold types and functions, incorporating both natural and engineered mutations while retaining conserved core functional regions. Details are provided in Appendix C.

3.2. Backbone Generation

We trained the SM and FM (with Optimal Transport) models on these protein families, using pretrained weights from Yim et al. (2023) and Bose et al. (2023), with each model having $\sim$ 17 million parameters. Each model generated 50 backbone structures, with target sequence lengths sampled from the distribution observed in the training data (Figure 5).

3.3. Backbone Dihedral Angles

In Figure 6, most data points fall within the allowed and favored regions, with few in the disallowed areas, and no significant geometric clashes or unreasonable conformations observed. The dihedral angle distributions of the generated structures align well with those of the experimentally derived proteins used for training. For instance, cytochrome c structures have sparse points in the region $-180° < \psi < -90°$ and $45° < \phi < 180°$, while GFP structures form four clusters there. SM samples are closer to the training data and concentrate in allowed regions and show less diversity than FM samples.

3.4. Conserved Residue Consistency

The optimal sequences preserved conserved residues that were largely consistent with the experimentally derived data (Figure 8). In Figure 7, similar to the experimentally derived proteins, the generated structures have increased residue conservation around binding pockets (refer to Figure 4 and Figure 12). For instance, in cytochrome c (1HRC) and generated SM-1 and FM-0, conserved residues cluster near the central heme C binding site. A similar pattern appears in 4OBE and the generated KRas proteins.

3.5. Structural Phylogenetic Tree

Summary tree. Pairwise structural distances between generated and experimental structures were computed using two methods, forming a matrix for phylogenetic tree construction. After normalizing branch lengths, a summary tree was generated with SumTrees (Moreno et al., 2024) (Figure 9). The topology shows that both the generated and the experimentally derived proteins cluster strongly by family, without intermixing among these groups.

Structure-informed tree outperforms sequence-only tree. Sequence-based phylogenetic trees were constructed by aligning sequences with Clustal Omega and inferring trees with FastTree (Price et al., 2009). Structure trees better preserve natural taxonomic groupings than sequence trees (Figure 10AB), especially at moderate to low sequence identity, and they achieve higher Taxonomic Congruence Scores (Tan et al., 2015) (Appendix G, Figure 10D), indicating closer alignment with the known taxonomy (Appendix G).

3.6. Dynamic Stability

Dynamic stability was analyzed as: (1) Backbone RMSD over time, with stable proteins typically below 2Å, or up to 3Å for larger, flexible proteins (Burton et al., 2012; Liu et al., 2017; Wong & Wong, 2024). In Figure 11A, experimentally derived structures remain within 2Å, while generated structures average $\sim$ 2.5Å, rarely exceeding 3Å. SM samples show RMSD $\sim$ 0.3Å higher than FM samples. (2) Radius of gyration (R_g) quantifies the spatial distribution of a molecule's atoms relative to its center of mass. Generated structures are expected to be compact, with R_g values close to experimentally derived structures and fluctuations around 1Å (Figure 11B). (3) The DSSP algorithm assigns secondary structure to each residue based on hydrogen bonds and geometry. In Figure 11D, the stability in secondary structure elements over time suggests structural stability with no major conformational changes. (4) Lower potential energy are generally more stable , with contributions from bond, angle, dihedral, van der Waals, and electrostatic energies. In Figure 11C, most generated samples have lower potential energy than the experimentally derived data.

3.7. Protein-ligand Docking

Docking on experimentally derived structures closely matches known binding modes, with ligand RMSDs $\sim$ 1.5Å and low binding energies (Figure 12). Successful docking typically shows binding free energies (ΔG) between -7 and -10 kcal/mol, with lower values indicate stronger, more stable interactions (Pushparathinam & Kathiravan, 2021; Nguyen et al., 2019). Generated structures also have pockets similar to those of experimentally derived proteins (Figure 12). In most simulations, ligands bind within 4Å of experimentally derived positions with ΔG below -6 kcal/mol. SM samples generally show lower ΔG and RMSDs than FM samples.

β-lactamases binding to penicillin. Mutations at Glu^{166} and Asn^{170} in class A β-lactamases (Figure 4A) can form a stable acyl-enzyme intermediate, disrupting deacylation (Chen & Herzberg, 2001). Only Asn and Gly at 170 preserve WT-like function (Brown et al., 2009), and this conservation

is retained in generated samples like FM-4 (Figure 12A), suggesting they may retain the ability to inactivate penicillin antibiotics. Notably, Asn/Gly at position 170 occurred in 23/50 FM samples and 18/50 SM samples, and when extended to positions 169 to 171, in 30/50 FM samples and 22/50 SM samples, rates higher than expected by chance.

Cytrochrome *c* binding to heme c. The generated cytochrome c-like structures retain conserved residues found in the WT (Figure 4B; 1HRC). SM-1 (Figure 12B) includes phosphorylatable residues Tyr^{58}, Thr^{59}, and Tyr^{107}, as well as lysine residues Lys^{82}, Lys^{83}, and Lys^{96} involved in phospholipid binding. Asn^{80} and these lysines form an ATP-binding pocket-like structure. FM-0 shows heme iron coordinated by two cysteines, which may form stronger covalent bonds, potentially affecting electron transfer efficiency.

KRas binding to GNP/GDP. GNP, a non-hydrolyzable GTP analog, is commonly used to simulate the GTP-bound active state of Ras proteins (Pantsar, 2020). In WT KRas bound to GNP (Figure 12D; 5UFE), the Switch II region residues fill the pocket. Removing of the γ-phosphate relaxes the conformation to the GDP-bound state, opening the switch II region (Figure 12C; 4OBE) (Kauke et al., 2017).

After blind docking the generated KRas-like structures with GNP/GDP, we analyzed the complexes using protein–ligand MD simulations. Allowing flexibility in both the backbone and side-chains enabled us to capture the dynamic conformational changes, especially the WT-like transitions in the switch regions, between active/inactive states. FM-generated structures show greater flexibility. In FM-13 (Figure 12D), GNP binding causes switch II residues to fill the pocket, similar to 5UFE. In the GDP-bound state, FM-13 adopts an open conformation, with the channel between the switches and P-loop opening, as seen in 4OBE (Figure 12C). In contrast, SM-27 are more rigid, with fewer conformational changes between GDP- and GTP-bound states.

4. Discussion

SM and FM can generate a range of monomeric protein structures and can have applications beyond protein design (Yim et al., 2023; Bose et al., 2023). However, the complexity of these tasks is often underestimated, and function verification for novel samples remains costly. Targeting specific tasks or integrating generative methods into well-established empirical knowledge may yield better results. Key factors like conformational dynamics, water interactions, and entropy have not been fully considered in generation (Du et al., 2024). Incorporating protein sequences, side-chain details, and functional annotations as context could improve model performance. Although some progress has been made (Torge et al., 2023; Jin et al., 2023; Somnath et al., 2023; Zhou et al., 2023), research gaps remain.

The generated samples for the most part capture the observed evolutionary diversity. However, there are regions of the observed diversity that are not so well covered (Figure 9) (such as GFP and class A β-lactamases), raising concerns about potential overfitting and limited generalizability. Common metrics are inadequate for quantifying overfitting in generative models (Arora et al., 2017), especially given the variability in sample sizes and quality across families.

5. Guidelines

Here, we outline a set of guidelines and best practices for designing and validating proteins using deep generative methods: (1) Implement physical constraints during the generation process to produce realistic structures without severe steric clashes. (2) Generate multiple candidate sequences for each design and prioritize those most likely to fold into the target structure. (3) Integrate sequence design and backbone validation early to ensure compatibility between the sequence and the backbone model. (4) Identify and preserve conserved residues in the designs to maintain critical functional motifs and overall structural integrity. (5) Construct all-atom models of the designs and refine them to relieve any strain or clashes introduced during initial modeling. (6) Compare the designs to known protein family structures or fold templates to ensure consistency with known protein folds. (7) Perform molecular dynamics simulations to verify that the proposed structures remain stable under physiologically relevant conditions. (8) Perform docking analyses to evaluate the viability of proposed ligand-binding sites. (9) Experimental validation is always essential to confirm the intended function and structural integrity of each candidate. (10) See Appendix J for circumstances under which implementation of these guidelines can be challenging.

6. Conclusion

This study shows the potential of methods based on deep generative models for designing proteins with high structural fidelity and functional plausibility. SM better captures conserved regions, producing more rigid structures, while FM offers greater flexibility (Figure 8). Although the optimized sequences for the generated structures exhibit some differences from natural family members, they preserve many functionally essential conserved residues (Figure 7). In structural phylogenetic trees (Figure 9), these designs cluster with their respective families, suggesting that their geometries could capture some family-specific signatures indicative of common ancestry or function. MD and docking simulations are consistent with the designs remain properly folded and stable under physiological conditions over time (Figure 11), and can accommodate their family-specific lig-

ands in binding pockets that closely correspond to known active sites (Figure 12). Some designs undergo conformational changes upon binding that mirror the allosteric behavior observed in WT (Figure 12CD).

Software and Data

All input data and software used in this study are available from public sources or provided under academic licenses. The source code, scripts, generated samples, and curated datasets can be accessed at https://github.com/ECburx/PROTEVAL. The PDB structure files were downloaded on June 19, 2024, from https://www.wwpdb.org/ftp/pdb-ftp-sites.

Acknowledgements

The authors would like to thank Charlie Harris, Simon Mathis and Stephen Eglen for their fruitful discussions and valuable feedback on this work.

References

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., and Lindahl, E. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1–2:19–25, September 2015. ISSN 2352-7110. doi: 10.1016/j.softx.2015.06.001. URL http://dx.doi.org/10.1016/j.softx.2015.06.001.

Ali, M., Ishqi, H. M., and Husain, Q. Enzyme engineering: Reshaping the biocatalytic functions. *Biotechnology and Bioengineering*, 117(6):1877–1894, March 2020. ISSN 1097-0290. doi: 10.1002/bit.27329. URL http://dx.doi.org/10.1002/bit.27329.

Anand, N. and Achim, T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models, 2022. URL https://arxiv.org/abs/2205.15019.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (gans), 2017. URL https://arxiv.org/abs/1703.00573.

Arunachalam, P. S., Walls, A. C., Golden, N., Atyeo, C., Fischinger, S., Li, C., Aye, P., Navarro, M. J., Lai, L., Edara, V. V., Röltgen, K., Rogers, K., Shirreff, L., Ferrell, D. E., Wrenn, S., Pettie, D., Kraft, J. C., Miranda, M. C., Kepl, E., Sydeman, C., Brunette, N., Murphy, M., Fiala, B., Carter, L., White, A. G., Trisal, M., Hsieh, C.-L., Russell-Lodrigue, K., Monjure, C., Dufour, J., Spencer, S., Doyle-Meyers, L., Bohm, R. P., Maness, N. J., Roy, C., Plante, J. A., Plante, K. S., Zhu, A., Gorman, M. J., Shin, S., Shen, X., Fontenot, J., Gupta, S., O'Hagan, D. T., Van Der Most, R., Rappuoli, R., Coffman, R. L., Novack, D., McLellan, J. S., Subramaniam, S., Montefiori, D., Boyd, S. D., Flynn, J. L., Alter, G., Villinger, F., Kleanthous, H., Rappaport, J., Suthar, M. S., King, N. P., Veesler, D., and Pulendran, B. Adjuvanting a subunit covid-19 vaccine to induce protective immunity. *Nature*, 594(7862):253–258, April 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03530-2. URL http://dx.doi.org/10.1038/s41586-021-03530-2.

Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., and Ben-Tal, N. Consurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Research*, 44 (W1):W344–W350, May 2016. ISSN 1362-4962. doi: 10.1093/nar/gkw408. URL http://dx.doi.org/10.1093/nar/gkw408.

Barclay, A. and Acharya, K. R. Engineering plastic eating enzymes using structural biology. *Biomolecules*, 13 (9):1407, September 2023. ISSN 2218-273X. doi: 10.3390/biom13091407. URL http://dx.doi.org/10.3390/biom13091407.

Barnard, T., Yu, X., Noinaj, N., and Taraska, J. Crystal structure of green fluorescent protein, April 2014. URL http://dx.doi.org/10.2210/pdb4KW4/pdb.

Behzadi, P., García-Perdomo, H. A., Karpiński, T. M., and Issakhanian, L. Metallo-β-lactamases: a review. *Molecular Biology Reports*, 47(8):6281–6294, July 2020. ISSN 1573-4978. doi: 10.1007/s11033-020-05651-9. URL http://dx.doi.org/10.1007/s11033-020-05651-9.

Berman, H. M. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, January 2000. ISSN 1362-4962. doi: 10.1093/nar/28.1.235. URL http://dx.doi.org/10.1093/nar/28.1.235.

Bose, A. J., Akhound-Sadegh, T., Huguet, G., Fatras, K., Rector-Brooks, J., Liu, C.-H., Nica, A. C., Korablyov, M., Bronstein, M., and Tong, A. SE(3)-stochastic flow matching for protein backbone generation, 2023. URL https://arxiv.org/abs/2310.02391.

Brown, N. G., Shanker, S., Prasad, B., and Palzkill, T. Structural and biochemical evidence that a tem-1 β-lactamase n170g active site mutant acts via substrate-assisted catalysis. *Journal of Biological Chemistry*, 284(48):33703–33712, November 2009. ISSN 0021-9258. doi: 10.1074/jbc.m109.053819. URL http://dx.doi.org/10.1074/jbc.M109.053819.

Burton, B., Zimmermann, M. T., Jernigan, R. L., and Wang, Y. A computational investigation on the con-

nection between dynamics properties of ribosomal proteins and ribosome assembly. *PLoS Computational Biology*, 8(5):e1002530, May 2012. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1002530. URL http://dx.doi.org/10.1371/journal.pcbi.1002530.

Bushnell, G. W., Louie, G. V., and Brayer, G. D. High-resolution three-dimensional structure of horse heart cytochrome c. *Journal of Molecular Biology*, 214 (2):585–595, July 1990. ISSN 0022-2836. doi: 10.1016/0022-2836(90)90200-6. URL http://dx.doi.org/10.1016/0022-2836(90)90200-6.

Chen, C. C. H. and Herzberg, O. Structures of the acyl-enzyme complexes of the staphylococcus aureus β-lactamase mutant glu166asp asn170gln with benzylpenicillin and cephaloridine. *Biochemistry*, 40(8):2351–2358, January 2001. ISSN 1520-4995. doi: 10.1021/bi002277h. URL http://dx.doi.org/10.1021/bi002277h.

Chen, R. T. Q. and Lipman, Y. Flow matching on general geometries. *ICLR 2024*, 2023. doi: 10.48550/ARXIV.2302.03660. URL https://arxiv.org/abs/2302.03660.

Cormack, B. P., Valdivia, R. H., and Falkow, S. Facs-optimized mutants of the green fluorescent protein (gfp). *Gene*, 173(1):33–38, 1996. ISSN 0378-1119. doi: 10.1016/0378-1119(95)00685-0. URL http://dx.doi.org/10.1016/0378-1119(95)00685-0.

Cox, A. Farnesyltransferase inhibitors: promises and realities. *Current Opinion in Pharmacology*, 2(4): 388–393, August 2002. ISSN 1471-4892. doi: 10.1016/s1471-4892(02)00181-9. URL http://dx.doi.org/10.1016/S1471-4892(02)00181-9.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., and Baker, D. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, October 2022. ISSN 1095-9203. doi: 10.1126/science.add2187. URL http://dx.doi.org/10.1126/science.add2187.

De Bortoli, V., Mathieu, E., Hutchinson, M., Thornton, J., Teh, Y. W., and Doucet, A. Riemannian score-based generative modelling, 2022. URL https://arxiv.org/abs/2202.02763.

Dickerson, R. E., Kopka, M. L., Borders, C. L., Varnum, J., Weinzierl, J. E., and Margoliash, E. A centrosymmetric projection at 4 Å of horse heart oxidized cytochrome c. *Journal of Molecular Biology*, 29(1):77–95, October 1967. ISSN 0022-2836. doi: 10.1016/0022-2836(67)90182-9. URL http://dx.doi.org/10.1016/0022-2836(67)90182-9.

Du, Y., Jamasb, A. R., Guo, J., Fu, T., Harris, C., Wang, Y., Duan, C., Liò, P., Schwaller, P., and Blundell, T. L. Machine learning-aided generative molecular design. *Nature Machine Intelligence*, 6(6): 589–604, June 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00843-5. URL http://dx.doi.org/10.1038/s42256-024-00843-5.

Eberhardt, J., Santos-Martins, D., Tillack, A. F., and Forli, S. Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, July 2021. ISSN 1549-960X. doi: 10.1021/acs.jcim.1c00203. URL http://dx.doi.org/10.1021/acs.jcim.1c00203.

Flores, T., Orengo, C., Moss, D., and Thornton, J. Comparison of conformational characteristics in structurally similar protein pairs. *Protein Science*, 2(11):1811–1826, November 1993. ISSN 1469-896X. doi: 10.1002/pro.5560021104. URL http://dx.doi.org/10.1002/pro.5560021104.

Glaser, A., McColl, B., and Vadolas, J. Gfp to bfp conversion: A versatile assay for the quantification of crispr/cas9-mediated genome editing. *Molecular Therapy - Nucleic Acids*, 5:e334, 2016. ISSN 2162-2531. doi: 10.1038/mtna.2016.48. URL http://dx.doi.org/10.1038/mtna.2016.48.

Greener, J. G., Moffat, L., and Jones, D. T. Design of metalloproteins and novel protein folds using variational autoencoders. *Scientific Reports*, 8(1), November 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-34533-1. URL http://dx.doi.org/10.1038/s41598-018-34533-1.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W., and Adams, P. D. Thecomputational crystallography toolbox: crystallographic algorithms in a reusable software framework. *Journal of Applied Crystallography*, 35(1):126–136, January 2002. ISSN 0021-8898. doi: 10.1107/s0021889801017824. URL http://dx.doi.org/10.1107/S0021889801017824.

Hollingsworth, S. A. and Dror, R. O. Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143, September 2018. ISSN 0896-6273. doi: 10.1016/j.neuron.2018.08.011. URL http://dx.doi.org/10.1016/j.neuron.2018.08.011.

Hooft, R. W. W., Vriend, G., Sander, C., and Abola, E. E. Errors in protein structures. *Nature*, 381(6580): 272–272, May 1996. ISSN 1476-4687. doi: 10.1038/381272a0. URL http://dx.doi.org/10.1038/381272a0.

Huang, P.-S., Boyken, S. E., and Baker, D. The coming of age of de novo protein design. *Nature*, 537(7620): 320–327, September 2016. ISSN 1476-4687. doi: 10.1038/nature19946. URL http://dx.doi.org/10.1038/nature19946.

Hunter, J. C., Gurbani, D., Ficarro, S. B., Carrasco, M. A., Lim, S. M., Choi, H. G., Xie, T., Marto, J. A., Chen, Z., Gray, N. S., and Westover, K. D. In situ selectivity profiling and crystal structure of sml-8-73-1, an active site inhibitor of oncogenic k-ras g12c. *Proceedings of the National Academy of Sciences*, 111 (24):8895–8900, June 2014. ISSN 1091-6490. doi: 10.1073/pnas.1404639111. URL http://dx.doi.org/10.1073/pnas.1404639111.

Hüttemann, M., Pecina, P., Rainbolt, M., Sanderson, T. H., Kagan, V. E., Samavati, L., Doan, J. W., and Lee, I. The multiple functions of cytochrome c and their regulation in life and death decisions of the mammalian cell: From respiration to apoptosis. *Mitochondrion*, 11(3):369–381, May 2011. ISSN 1567-7249. doi: 10.1016/j.mito.2011.01.010. URL http://dx.doi.org/10.1016/j.mito.2011.01.010.

Illergård, K., Ardell, D. H., and Elofsson, A. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, 77(3):499–508, June 2009. ISSN 1097-0134. doi: 10.1002/prot.22458. URL http://dx.doi.org/10.1002/prot.22458.

Ingraham, J. B., Baranov, M., Costello, Z., Barber, K. W., Wang, W., Ismail, A., Frappier, V., Lord, D. M., Ng-Thow-Hing, C., Van Vlack, E. R., Tie, S., Xue, V., Cowles, S. C., Leung, A., Rodrigues, J. V., Morales-Perez, C. L., Ayoub, A. M., Green, R., Puentes, K., Oplinger, F., Panwar, N. V., Obermeyer, F., Root, A. R., Beam, A. L., Poelwijk, F. J., and Grigoryan, G. Illuminating protein space with a programmable generative model. *Nature*, 623(7989): 1070–1078, November 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06728-8. URL http://dx.doi.org/10.1038/s41586-023-06728-8.

Jamasb, A. R., Morehead, A., Joshi, C. K., Zhang, Z., Didi, K., Mathis, S. V., Harris, C., Tang, J., Cheng, J., Lio, P., and Blundell, T. L. Evaluating representation learning on the protein structure universe, 2024. URL https://arxiv.org/abs/2406.13864.

Jelsch, C., Mourey, L., Masson, J., and Samama, J. Crystal structure of escherichia coli tem1 β-lactamase at 1.8 Å resolution. *Proteins: Structure, Function, and Bioinformatics*, 16(4):364–383, August 1993. ISSN 1097-0134. doi: 10.1002/prot.340160406. URL http://dx.doi.org/10.1002/prot.340160406.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023. ISSN 1557-7341. doi: 10.1145/3571730. URL http://dx.doi.org/10.1145/3571730.

Jin, W., Sarkizova, S., Chen, X., Hacohen, N., and Uhler, C. Unsupervised protein-ligand binding energy prediction via neural euler's rotation equation, 2023. URL https://arxiv.org/abs/2301.10814.

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., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596 (7873):583–589, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL http://dx.doi.org/10.1038/s41586-021-03819-2.

Kagan, V. E., Bayır, H. A., Belikova, N. A., Kapralov, O., Tyurina, Y. Y., Tyurin, V. A., Jiang, J., Stoyanovsky, D. A., Wipf, P., Kochanek, P. M., Greenberger, J. S., Pitt, B., Shvedova, A. A., and Borisenko, G. Cytochrome c/cardiolipin relations in mitochondria: a kiss of death. *Free Radical Biology and Medicine*, 46(11):1439–1453, June 2009. ISSN 0891-5849. doi: 10.1016/j.freeradbiomed.2009.03.004. URL http://dx.doi.org/10.1016/j.freeradbiomed.2009.03.004.

Kashyap, D., Garg, V. K., and Goel, N. *Intrinsic and extrinsic pathways of apoptosis: Role in cancer development and prognosis*, pp. 73–120. Elsevier, 2021. ISBN 9780323853156. doi: 10.1016/bs.apcsb.2021.01.003. URL http://dx.doi.org/10.1016/bs.apcsb.2021.01.003.

Kauke, M. J., Traxlmayr, M. W., Parker, J. A., Kiefer, J. D., Knihtila, R., McGee, J., Verdine, G., Mattos, C., and Wittrup, K. D. An engineered protein antagonist of k-ras/b-raf interaction. *Scientific Reports*, 7(1), July 2017. ISSN 2045-2322. doi: 10.1038/

s41598-017-05889-7. URL http://dx.doi.org/10.1038/s41598-017-05889-7.

Kingwell, K. Ribosome-targeted antibiotic tackles antimicrobial resistance. *Nature Reviews Drug Discovery*, 23(4):249–249, March 2024. ISSN 1474-1784. doi: 10.1038/d41573-024-00040-4. URL http://dx.doi.org/10.1038/d41573-024-00040-4.

Korendovych, I. V. and DeGrado, W. F. De novo protein design, a retrospective. *Quarterly Reviews of Biophysics*, 53, 2020. ISSN 1469-8994. doi: 10.1017/s0033583519000131. URL http://dx.doi.org/10.1017/S0033583519000131.

Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G. R., Morey-Burrows, F. S., Anishchenko, I., Humphreys, I. R., McHugh, R., Vafeados, D., Li, X., Sutherland, G. A., Hitchcock, A., Hunter, C. N., Kang, A., Brackenbrough, E., Bera, A. K., Baek, M., DiMaio, F., and Baker, D. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, 384(6693), April 2024. ISSN 1095-9203. doi: 10.1126/science.adl2528. URL http://dx.doi.org/10.1126/science.adl2528.

Krissinel, E. and Henrick, K. Secondary-structure matching (ssm), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D Biological Crystallography*, 60(12):2256–2268, November 2004. ISSN 0907-4449. doi: 10.1107/s0907444904026460. URL http://dx.doi.org/10.1107/S0907444904026460.

Ladygina, N., Martin, B. R., and Altman, A. *Dynamic Palmitoylation and the Role of DHHC Proteins in T Cell Activation and Anergy*, pp. 1–44. Elsevier, 2011. ISBN 9780123876645. doi: 10.1016/b978-0-12-387664-5.00001-7. URL http://dx.doi.org/10.1016/B978-0-12-387664-5.00001-7.

Laskowski, R., Rullmann, J., MacArthur, M., Kaptein, R., and Thornton, J. Aqua and procheck-nmr: Programs for checking the quality of protein structures solved by nmr. *Journal of Biomolecular NMR*, 8(4), December 1996. ISSN 1573-5001. doi: 10.1007/bf00228148. URL http://dx.doi.org/10.1007/BF00228148.

Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. Procheck: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2):283–291, April 1993. ISSN 0021-8898. doi: 10.1107/s0021889892009944. URL http://dx.doi.org/10.1107/S0021889892009944.

Lee, D., Das, S., Dawson, N. L., Dobrijevic, D., Ward, J., and Orengo, C. Novel computational protocols for functionally classifying and characterising serine beta-lactamases. *PLOS Computational Biology*, 12(6):e1004926, June 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1004926. URL http://dx.doi.org/10.1371/journal.pcbi.1004926.

Lexa, K. W. and Carlson, H. A. Protein flexibility in docking and surface mapping. *Quarterly Reviews of Biophysics*, 45(3):301–343, May 2012. ISSN 1469-8994. doi: 10.1017/s0033583512000066. URL http://dx.doi.org/10.1017/S0033583512000066.

Li, M., Wu, X., and Xu, X.-C. Induction of apoptosis by cyclo-oxygenase-2 inhibitor ns398 through a cytochromec-dependent pathway in esophageal cancer cells. *International Journal of Cancer*, 93(2):218–223, 2001. ISSN 1097-0215. doi: 10.1002/ijc.1322. URL http://dx.doi.org/10.1002/ijc.1322.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling, 2022. URL https://arxiv.org/abs/2210.02747.

Liu, K., Watanabe, E., and Kokubo, H. Exploring the stability of ligand binding modes to proteins by molecular dynamics simulations. *Journal of Computer-Aided Molecular Design*, 31(2):201–211, January 2017. ISSN 1573-4951. doi: 10.1007/s10822-016-0005-2. URL http://dx.doi.org/10.1007/s10822-016-0005-2.

Malik, A. J., Poole, A. M., and Allison, J. R. Structural phylogenetics with confidence. *Molecular Biology and Evolution*, 37(9):2711–2726, April 2020. ISSN 1537-1719. doi: 10.1093/molbev/msaa100. URL http://dx.doi.org/10.1093/molbev/msaa100.

Mayrose, I. Comparison of site-specific rate-inference methods for protein sequences: Empirical bayesian methods are superior. *Molecular Biology and Evolution*, 21(9):1781–1791, May 2004. ISSN 1537-1719. doi: 10.1093/molbev/msh194. URL http://dx.doi.org/10.1093/molbev/msh194.

McIntosh, D. B., Parrish, J. C., and Wallace, C. J. Definition of a nucleotide binding site on cytochrome c by photoaffinity labeling. *Journal of Biological Chemistry*, 271(31):18379–18386, August 1996. ISSN 0021-9258. doi: 10.1074/jbc.271.31.18379. URL http://dx.doi.org/10.1074/jbc.271.31.18379.

Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., and Dessimoz, C. Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. September 2023. doi: 10.1101/2023.09.19.558401.

URL http://dx.doi.org/10.1101/2023.09.19.558401.

Moreno, M. A., Holder, M. T., and Sukumaran, J. Dendropy 5: a mature python library for phylogenetic computing, 2024. URL https://doi.org/10.21105/joss.06943.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., and Olson, A. J. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16):2785–2791, April 2009. ISSN 1096-987X. doi: 10.1002/jcc.21256. URL http://dx.doi.org/10.1002/jcc.21256.

Naas, T., Oueslati, S., Bonnin, R. A., Dabos, M. L., Zavala, A., Dortet, L., Retailleau, P., and Iorga, B. I. Beta-lactamase database (bldb) – structure and function. *Journal of Enzyme Inhibition and Medicinal Chemistry*, 32 (1):917–919, January 2017. ISSN 1475-6374. doi: 10.1080/14756366.2017.1344235. URL http://dx.doi.org/10.1080/14756366.2017.1344235.

Nguyen, N. T., Nguyen, T. H., Pham, T. N. H., Huy, N. T., Bay, M. V., Pham, M. Q., Nam, P. C., Vu, V. V., and Ngo, S. T. Autodock vina adopts more accurate binding poses but autodock4 forms better binding affinity. *Journal of Chemical Information and Modeling*, 60(1):204–211, December 2019. ISSN 1549-960X. doi: 10.1021/acs.jcim.9b00778. URL http://dx.doi.org/10.1021/acs.jcim.9b00778.

Ormö, M., Cubitt, A. B., Kallio, K., Gross, L. A., Tsien, R. Y., and Remington, S. J. Crystal structure of the aequorea victoria green fluorescent protein. *Science*, 273 (5280):1392–1395, September 1996. ISSN 1095-9203. doi: 10.1126/science.273.5280.1392. URL http://dx.doi.org/10.1126/science.273.5280.1392.

Ow, Y.-L. P., Green, D. R., Hao, Z., and Mak, T. W. Cytochrome c: functions beyond respiration. *Nature Reviews Molecular Cell Biology*, 9(7):532–542, July 2008. ISSN 1471-0080. doi: 10.1038/nrm2434. URL http://dx.doi.org/10.1038/nrm2434.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1), October 2011. ISSN 1758-2946. doi: 10.1186/1758-2946-3-33. URL http://dx.doi.org/10.1186/1758-2946-3-33.

Pantsar, T. The current understanding of kras protein structure and dynamics. *Computational and Structural Biotechnology Journal*, 18:189–198, 2020. ISSN 2001-0370. doi: 10.1016/j.csbj.2019.12.004. URL http://dx.doi.org/10.1016/j.csbj.2019.12.004.

Patient, S., Wieser, D., Kleen, M., Kretschmann, E., Jesus Martin, M., and Apweiler, R. Uniprotjapi: a remote api for accessing uniprot data. *Bioinformatics*, 24(10):1321–1322, April 2008. ISSN 1367-4803. doi: 10.1093/bioinformatics/btn122. URL http://dx.doi.org/10.1093/bioinformatics/btn122.

Pollard, D. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, December 2001. ISBN 9780511811555. doi: 10.1017/cbo9780511811555. URL http://dx.doi.org/10.1017/CBO9780511811555.

Pooladian, A.-A., Ben-Hamu, H., Domingo-Enrich, C., Amos, B., Lipman, Y., and Chen, R. T. Q. Multisample flow matching: Straightening flows with minibatch couplings, 2023. URL https://arxiv.org/abs/2304.14772.

Price, M. N., Dehal, P. S., and Arkin, A. P. Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26(7):1641–1650, April 2009. ISSN 1537-1719. doi: 10.1093/molbev/msp077. URL http://dx.doi.org/10.1093/molbev/msp077.

Puente-Lelievre, C., Malik, A. J., Douglas, J., Ascher, D., Baker, M., Allison, J., Poole, A., Lundin, D., Fullmer, M., Bouckert, R., Kim, H., Steinegger, M., and Matzke, N. Tertiary-interaction characters enable fast, model-based structural phylogenetics beyond the twilight zone. December 2023. doi: 10.1101/2023.12.12.571181. URL http://dx.doi.org/10.1101/2023.12.12.571181.

Pupko, T., Bell, R. E., Mayrose, I., Glaser, F., and Ben-Tal, N. Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, 18(suppl_1):S71–S77, July 2002. ISSN 1367-4803. doi: 10.1093/bioinformatics/18.suppl_1.s71. URL http://dx.doi.org/10.1093/bioinformatics/18.suppl_1.S71.

Pushparathinam, G. and Kathiravan, M. K. Docking studies and molecular dynamics simulation of triazole benzene sulfonamide derivatives with human carbonic anhydrase ix inhibition activity. *RSC Advances*, 11 (60):38079–38093, 2021. ISSN 2046-2069. doi: 10.1039/d1ra07377j. URL http://dx.doi.org/10.1039/D1RA07377J.

Pédelacq, J.-D., Cabantous, S., Tran, T., Terwilliger, T. C., and Waldo, G. S. Engineering and characterization of a superfolder green fluorescent protein. *Nature Biotechnology*, 24(1):79–88, December 2005. ISSN 1546-1696.

doi: 10.1038/nbt1172. URL http://dx.doi.org/10.1038/nbt1172.

Ramachandran, G., Ramakrishnan, C., and Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95–99, July 1963. ISSN 0022-2836. doi: 10.1016/s0022-2836(63)80023-6. URL http://dx.doi.org/10.1016/S0022-2836(63)80023-6.

Remington, S. J. Green fluorescent protein: A perspective. *Protein Science*, 20(9):1509–1519, July 2011. ISSN 1469-896X. doi: 10.1002/pro.684. URL http://dx.doi.org/10.1002/pro.684.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 2019. doi: 10.1101/622803. URL https://www.biorxiv.org/content/10.1101/622803v4.

Rudolph, M., Wandt, B., and Rosenhahn, B. Same same but differnet: Semi-supervised defect detection with normalizing flows, 2020. URL https://arxiv.org/abs/2008.12577.

Salemme, F. R. Structure and function of cytochromes c. *Annual Review of Biochemistry*, 46(1):299–330, June 1977. ISSN 1545-4509. doi: 10.1146/annurev.bi.46.070177.001503. URL http://dx.doi.org/10.1146/annurev.bi.46.070177.001503.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology*, 7(1), January 2011. ISSN 1744-4292. doi: 10.1038/msb.2011.75. URL http://dx.doi.org/10.1038/msb.2011.75.

Simanshu, D. K., Nissley, D. V., and McCormick, F. Ras proteins and their regulators in human disease. *Cell*, 170(1):17–33, June 2017. ISSN 0092-8674. doi: 10.1016/j.cell.2017.06.009. URL http://dx.doi.org/10.1016/j.cell.2017.06.009.

Singh, S. B. and Lingham, R. B. Current progress on farnesyl protein transferase inhibitors. *Current Opinion in Drug Discovery and Development*, 5(2):225 – 244, 2002.

Somnath, V. R., Pariset, M., Hsieh, Y.-P., Martinez, M. R., Krause, A., and Bunne, C. Aligned diffusion schrödinger bridges, 2023. URL https://arxiv.org/abs/2302.11419.

Soteras Gutiérrez, I., Lin, F.-Y., Vanommeslaeghe, K., Lemkul, J. A., Armacost, K. A., Brooks, C. L., and MacKerell, A. D. Parametrization of halogen bonds in the charmm general force field: Improved treatment of ligand–protein interactions. *Bioorganic & Medicinal Chemistry*, 24(20):4812–4825, October 2016. ISSN 0968-0896. doi: 10.1016/j.bmc.2016.06.034. URL http://dx.doi.org/10.1016/j.bmc.2016.06.034.

Tan, G., Gil, M., Löytynoja, A. P., Goldman, N., and Dessimoz, C. Simple chained guide trees give poorer multiple sequence alignments than inferred trees in simulation and phylogenetic benchmarks. *Proceedings of the National Academy of Sciences*, 112(2), January 2015. ISSN 1091-6490. doi: 10.1073/pnas.1417526112. URL http://dx.doi.org/10.1073/pnas.1417526112.

Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., and Bengio, Y. Improving and generalizing flow-based generative models with minibatch optimal transport, 2023. URL https://arxiv.org/abs/2302.00482.

Tooke, C. L., Hinchliffe, P., Bragginton, E. C., Colenso, C. K., Hirvonen, V. H., Takebayashi, Y., and Spencer, J. β-lactamases and β-lactamase inhibitors in the 21st century. *Journal of Molecular Biology*, 431(18):3472–3500, August 2019. ISSN 0022-2836. doi: 10.1016/j.jmb.2019.04.002. URL http://dx.doi.org/10.1016/j.jmb.2019.04.002.

Torge, J., Harris, C., Mathis, S. V., and Lio, P. Diffhopp: A graph diffusion model for novel drug design via scaffold hopping, 2023. URL https://arxiv.org/abs/2308.07416.

Trott, O. and Olson, A. J. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, June 2009. ISSN 1096-987X. doi: 10.1002/jcc.21334. URL http://dx.doi.org/10.1002/jcc.21334.

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., and Steinegger, M. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42(2):243–246, May 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0. URL http://dx.doi.org/10.1038/s41587-023-01773-0.

Vanommeslaeghe, K. and MacKerell, A. D. Automation of the charmm general force field (cgenff) i: Bond perception and atom typing. *Journal of Chemical Information and Modeling*, 52(12):3144–3154, November 2012. ISSN 1549-960X. doi: 10.1021/ci300363c. URL http://dx.doi.org/10.1021/ci300363c.

Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., and Mackerell, A. D. Charmm general force field: A force field for drug-like molecules compatible with the charmm all-atom additive biological force fields. *Journal of Computational Chemistry*, 31(4):671–690, July 2009. ISSN 1096-987X. doi: 10.1002/jcc.21367. URL http://dx.doi.org/10.1002/jcc.21367.

Vanommeslaeghe, K., Raman, E. P., and MacKerell, A. D. Automation of the charmm general force field (cgenff) ii: Assignment of bonded parameters and partial atomic charges. *Journal of Chemical Information and Modeling*, 52(12):3155–3168, November 2012. ISSN 1549-960X. doi: 10.1021/ci3003649. URL http://dx.doi.org/10.1021/ci3003649.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

Victorino da Silva Amatto, I., Gonsales da Rosa-Garzon, N., Antônio de Oliveira Simões, F., Santiago, F., Pereira da Silva Leite, N., Raspante Martins, J., and Cabral, H. Enzyme engineering and its industrial applications. *Biotechnology and Applied Biochemistry*, 69(2):389–409, February 2021. ISSN 1470-8744. doi: 10.1002/bab.2117. URL http://dx.doi.org/10.1002/bab.2117.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7): 1661–1674, July 2011. ISSN 1530-888X. doi: 10.1162/neco_a_00142. URL http://dx.doi.org/10.1162/NECO_a_00142.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., De Bortoli, V., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL http://dx.doi.org/10.1038/s41586-023-06415-8.

Way, T.-D., Kao, M.-C., and Lin, J.-K. Degradation of her2/neu by apigenin induces apoptosis through cytochrome c release and caspase-3 activation in her2/neu-overexpressing breast cancer cells. *FEBS Letters*, 579 (1):145–152, November 2004. ISSN 1873-3468. doi: 10.1016/j.febslet.2004.11.061. URL http://dx.doi.org/10.1016/j.febslet.2004.11.061.

Weinmann, H. and Ottow, E. *Recent Development in Novel Anticancer Therapies*, pp. 221–251. Elsevier, 2007. ISBN 9780080450445. doi: 10.1016/b0-08-045044-x/00210-8. URL http://dx.doi.org/10.1016/B0-08-045044-X/00210-8.

Winnifrith, A., Outeiral, C., and Hie, B. L. Generative artificial intelligence for de novo protein design. *Current Opinion in Structural Biology*, 86:102794, June 2024. ISSN 0959-440X. doi: 10.1016/j.sbi.2024.102794. URL http://dx.doi.org/10.1016/j.sbi.2024.102794.

Wong, H. Y. and Wong, K.-B. *Using AlphaFold2 and Molecular Dynamics Simulation to Model Protein Recognition*, pp. 49–66. Springer US, 2024. ISBN 9781071640593. doi: 10.1007/978-1-0716-4059-3_4. URL http://dx.doi.org/10.1007/978-1-0716-4059-3_4.

Wu, K. E., Yang, K. K., Berg, R. v. d., Zou, J. Y., Lu, A. X., and Amini, A. P. Protein structure generation via folding diffusion, 2022. URL https://arxiv.org/abs/2209.15611.

Yariv, B., Yariv, E., Kessel, A., Masrati, G., Chorin, A. B., Martz, E., Mayrose, I., Pupko, T., and Ben-Tal, N. Using evolutionary data to make sense of macromolecules with a "face-lifted" consurf. *Protein Science*, 32(3), February 2023. ISSN 1469-896X. doi: 10.1002/pro.4582. URL http://dx.doi.org/10.1002/pro.4582.

Yim, J., Trippe, B. L., De Bortoli, V., Mathieu, E., Doucet, A., Barzilay, R., and Jaakkola, T. SE(3) diffusion model with application to protein backbone generation. *ICML*, 2023. doi: 10.48550/ARXIV.2302.02277. URL https://arxiv.org/abs/2302.02277.

Yu, W., He, X., Vanommeslaeghe, K., and MacKerell, A. D. Extension of the charmm general force field to sulfonyl-containing compounds and its utility in biomolecular simulations. *Journal of Computational Chemistry*, 33 (31):2451–2468, July 2012. ISSN 1096-987X. doi: 10.1002/jcc.23067. URL http://dx.doi.org/10.1002/jcc.23067.

Zhang, Y. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic Acids Research*, 33(7): 2302–2309, April 2005. ISSN 1362-4962. doi: 10.1093/nar/gki524. URL http://dx.doi.org/10.1093/nar/gki524.

Zhou, B., Zheng, L., Wu, B., Yi, K., Zhong, B., Tan, Y., Liu, Q., Liò, P., and Hong, L. A conditional protein diffusion model generates artificial programmable endonuclease sequences with enhanced activity. August 2023. doi: 10.1101/2023.08.10.552783. URL http://dx.doi.org/10.1101/2023.08.10.552783.

Šali, A. and Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, December 1993. ISSN 0022-2836. doi: 10.1006/jmbi.1993.1626. URL http://dx.doi.org/10.1006/jmbi.1993.1626.

A. Deep Generative Protein Design Workflow

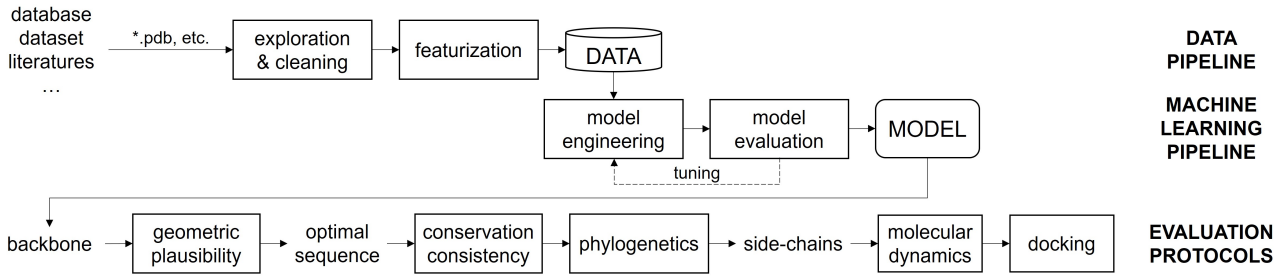

Figure 1. A schematic overview of the deep generative protein design pipeline and its evaluation protocols.

B. Protein Backbone Generation

B.1. $\mathrm{SE}(3)$ Decomposition into $\mathrm{SO}(3)$ and $\mathbb{R}^3$

The Special Euclidean group $\mathrm{SE}(3)$ describes the rotations and translations in 3D space. An element of $\mathrm{SE}(3)$ can be represented by a 4×4 matrix:

$$\mathbf{T} = \begin{pmatrix} \mathbf{R} & \mathbf{x} \\ \mathbf{0}_{1\times 3} & 1 \end{pmatrix} \tag{1}$$

where $\mathbf{R}$ is a 3×3 rotation matrix belonging to the Special Orthogonal group $\mathrm{SO}(3)$, and $\mathbf{x} = [\mathrm{x}_x\ \mathrm{x}_y\ \mathrm{x}_z] \in \mathbb{R}^3$ is the translational component. Since $\mathrm{SE}(3)$ can be viewed as the semidirect product of $\mathrm{SO}(3)$ and $\mathbb{R}^3$, denoted as $\mathrm{SE}(3) \cong \mathrm{SO}(3) \ltimes \mathbb{R}^3$, one option is to naturally treat $\mathrm{SO}(3)$ and $\mathbb{R}^3$ as independent for simplicity (Yim et al., 2023).

B.2. Protein Backbone Representations

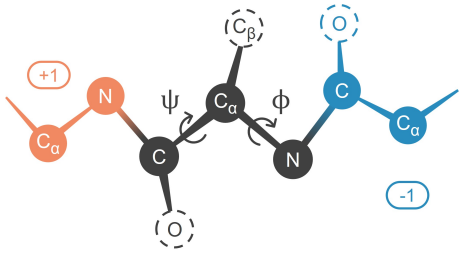

Figure 2. Protein backbone with dihedral angles ψ and ϕ.

Molecules can be intuitively represented as 3D atomic point clouds. However, macromolecules like proteins may contain thousands or tens of thousands of atoms, with variation in the atom types and quantities among different amino acids (for instance, sulfur atoms are present only in a few amino acids like cysteine). Representing proteins as unordered 3D atomic point clouds significantly increases data dimensionality and sparsity, requiring far more training data than is typically available.

Following the work of Yim et al. (2023) and Bose et al. (2023), we adopt the more compact backbone rigid groups from AlphaFold (Jumper et al., 2021) to represent protein backbone structures in 3D space. A backbone rigid group consists of the main chain atoms (N, C_α, C, O) within a single residue (Figure 2), where their geometric relationships (relative positions and orientations) are highly stable. The position and orientation of the group is transformed as a whole, without accounting for individual atomic movements, simplifying the computation and reducing structural errors caused by excessive degrees of freedom.

Assuming experimentally derived ideal chemical bond angles and lengths, models learn how the rigid transformation (or *frame*) $\mathbf{T}_i$ of each residue $i \in [1, N]$ acts on idealized coordinates $[\mathrm{N}^\star, \mathrm{C}_\alpha^\star, \mathrm{C}^\star] \in \mathbb{R}^3$ (centered at $\mathrm{C}_\alpha^\star$), so that the transformed coordinates match the actual coordinates as closely as possible:

$$[\mathrm{N}, \mathrm{C}_\alpha, \mathrm{C}]_i = \mathbf{T}_i \cdot [\mathrm{N}^\star, \mathrm{C}_\alpha^\star, \mathrm{C}^\star] \tag{2}$$

where $\mathbf{T}_i \in \mathrm{SE}(3)$ can be decomposed into a rotation matrix $\mathbf{R}_i \in \mathrm{SO}(3)$ and a translation vector $\mathbf{x}_i \in \mathbb{R}^3$. An additional torsion angle $\psi_i \in \mathrm{SO}(2)$ is introduced between the bond of C_α and C for a more accurate construction of the backbone oxygen atom O.

B.3. $\mathrm{SE}(3)$ Score Matching

Let $\mathbf{T}_t = [\mathbf{T}_{1,t}, \ldots, \mathbf{T}_{N,t}] \in \mathrm{SE}(3)^N$ denote the manifold of N frames at time t, where each frame can independently rotate and translate. Correspondingly, define $\mathbf{R}_t = [\mathbf{R}_{1,t}, \ldots, \mathbf{R}_{N,t}]$ and $\mathbf{X}_t = [\mathbf{x}_{1,t}, \ldots, \mathbf{x}_{N,t}]$. By treating $\mathrm{SO}(3)$ and $\mathbb{R}^3$ as two independent stochastic processes, a *forward process* gradually perturbs the initial data distribution p_0. Following the approach of Yim et al. (2023), this process is described by the Stochastic Differential Equation (SDE) for $\mathbf{T}_t \sim p_t$ and any arbitrary time $t \in [0, T]$:

$$\mathrm{d}\mathbf{T}_t = \left[0, -\frac{1}{2}\mathbf{X}_t\right] \mathrm{d}t + \left[\mathrm{d}\mathbf{B}_t^{\mathrm{SO}(3)}, \mathrm{d}\mathbf{B}_t^{\mathbb{R}^3}\right] \tag{3}$$

where $\mathbf{B}_t^{\mathrm{SO}(3)}$ and $\mathbf{B}_t^{\mathbb{R}^3}$ are Brownian motions on the $\mathrm{SO}(3)$ and $\mathbb{R}^3$, respectively. Invariant density $p_{\mathrm{inv}}^{\mathrm{SE}(3)}(\mathbf{T}) \propto \mathcal{U}^{\mathrm{SO}(3)}(\mathbf{R})\,\mathcal{N}(\mathbf{x}; 0, \mathbf{I})$ is chosen for $\mathbf{T} = (\mathbf{R}, \mathbf{x})$.

Let $(\overleftarrow{\mathbf{T}}_t)_{t\in[0,T]} = (\mathbf{T}_{T-t})_{t\in[0,T]}$, the corresponding *time-reverse process* (De Bortoli et al., 2022) is given by

$$\mathrm{d}\overleftarrow{\mathbf{R}}_t = \nabla_{\mathbf{R}} \log p_{T-t}(\overleftarrow{\mathbf{T}}_t)\mathrm{d}t + \mathrm{d}\mathbf{B}_t^{\mathrm{SO}(3)} \tag{4}$$

$$\mathrm{d}\overleftarrow{\mathbf{X}}_t = \left\{ \frac{\overleftarrow{\mathbf{X}}_t}{2} + \nabla_{\mathbf{x}} \log p_{T-t}(\overleftarrow{\mathbf{T}}_t) \right\} \mathrm{d}t + \mathrm{d}\mathbf{B}_t^{\mathbb{R}^3} \tag{5}$$

where $\nabla \log p$ is the gradient of the log-probability density function (also known as the Stein *score*). However, this gradient is typically intractable in practice because the exact form of $p_t(\mathbf{T}_t)$ is unknown at any given time t.

Instead, score-based models estimate tractable conditional score $\nabla \log p_{t|0}$ through SM (Vincent, 2011), using a neural network $s(\theta, t, \cdot)$ trained by minimizing both[3]:

$$\mathcal{L}_{\mathrm{SM}}^{\mathbf{R}}(\theta) = \mathbb{E}\left[\|\nabla_{\mathbf{R}} \log p_{t|0}(\mathbf{R}_t|\mathbf{R}_0) - s(\theta, t, \mathbf{R}_t)\|^2\right] \tag{6}$$

$$\mathcal{L}_{\mathrm{SM}}^{\mathbf{X}}(\theta) = \mathbb{E}\left[\|\nabla_{\mathbf{X}} \log p_{t|0}(\mathbf{X}_t|\mathbf{X}_0) - s(\theta, t, \mathbf{X}_t)\|^2\right] \tag{7}$$

with $t \sim \mathcal{U}(0, T)$ and

$$\nabla_{\mathbf{R}} \log p_{t|0}(\mathbf{R}_t|\mathbf{R}_0) = \frac{\mathbf{R}_t}{\omega(\mathbf{R}_{0\to t})} \log\{\mathbf{R}_{0\to t}\} \frac{\partial_\omega f(\omega(\mathbf{R}_{0\to t}), t)}{f(\omega(\mathbf{R}_{0\to t}), t)} \tag{8}$$

$$\nabla_{\mathbf{x}} \log p_{t|0}(\mathbf{x}_t|\mathbf{x}_0) = \frac{e^{-t/2}\mathbf{x}_0 - \mathbf{x}_t}{1 - e^{-t}} \tag{9}$$

where ω represents the rotation angle, $\mathbf{R}_{0\to t} = \mathbf{R}_0^\top \mathbf{R}_t$, and

$$f(\omega, t) = \sum_{\ell\in\mathbb{N}} (2\ell + 1) e^{-\ell(\ell+1)t/2} \frac{\sin((\ell + \frac{1}{2})\omega)}{\sin(\frac{\omega}{2})} \tag{10}$$

is an auxiliary function for the heat kernel[4] of the Brownian motion on $\mathrm{SO}(3)$.

B.4. $\mathrm{SE}(3)$ Flow Matching

FM is a simulation-free method for training vector fields to follow a prescribed conditional probability path (Lipman et al., 2022). Formally, for $t \in [0, 1]$, let $\mathbf{U} = \{\mathbf{u}_t\}$ be a *flow* which is a set of time-indexed vector fields that describe the paths along which data points move from an initial distribution p_1 to a target distribution p_0. Each vector field $\mathbf{u}_t(\mathbf{T}_t)$ represents the rate of change of $\mathbf{T}_t$ which is typically the solution to the Ordinary Differential Equation (ODE) $\frac{d}{dt}\mathbf{T}_t = \mathbf{u}_t(\mathbf{T}_t)$. FM approximates $\mathbf{u}_t(\mathbf{T}_t)$ with a network $v(\theta, t, \cdot)$ by minimizing $\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}\,\|\mathbf{u}_t(\mathbf{T}_t) - v(\theta, t, \mathbf{T}_t)\|^2$ with $t \sim \mathcal{U}(0, 1)$.

Similarly, independent flows can be built on $\mathrm{SO}(3)$ and $\mathbb{R}^3$. Computing $\mathbf{u}_t$, however, is also intractable due to the complex integrals involved in defining the marginal probability path and vector field. By showing $\nabla_\theta \mathcal{L}_{\mathrm{FM}}(\theta) = \nabla_\theta \mathcal{L}_{\mathrm{CFM}}(\theta)$, Lipman et al. (2022) suggested the tractable conditional FM objective on $\mathbb{R}^3$:

$$\mathcal{L}_{\mathrm{CFM}}^{\mathbf{X}}(\theta) = \mathbb{E}\,\|\mathbf{u}_t(\mathbf{X}_t|\mathbf{X}_0) - v(\theta, t, \mathbf{X}_t)\|^2 \tag{11}$$

with the Gaussian path $p_t(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; t\mathbf{x}_0, (t\sigma - t + 1)^2)$ generated by

$$\mathbf{u}_t(\mathbf{x}_t|\mathbf{x}_0) = \frac{\mathbf{x}_0 - (1 - \sigma)\mathbf{x}_t}{1 - (1 - \sigma)t} \tag{12}$$

[3] One adds weights $1/\mathbb{E}[\|\nabla_{\mathbf{R}} \log p_{t|0}(\mathbf{R}_t|\mathbf{R}_0)\|^2]$ to Equation (6) and $(1 - e^{-t})/e^{-t/2}$ to Equation (7) for simplicity.

[4] The heat kernel on a manifold is the fundamental solution to the heat equation, representing the probability density function of a Brownian particle diffusing from one point to another over time.

where $\sigma > 0$ is a smoothing constant.

For flows on $\mathrm{SO}(3)$, Bose et al. (2023) set

$$\mathcal{L}_{\mathrm{CFM}}^{\mathbf{R}}(\theta) = \mathbb{E}\left\|\mathbf{u}_t(\mathbf{R}_t|\mathbf{R}_0,\mathbf{R}_1) - \mathbf{v}(\theta,t,\mathbf{R}_t)\right\|^2 \tag{13}$$

and define the geodesic interpolant between $\mathbf{R}_1 \sim p_1$ and $\mathbf{R}_0 \sim p_0$ as $\mathbf{R}_t = \exp_{\mathbf{R}_1}(t\log_{\mathbf{R}_1}(\mathbf{R}_0))$. Let Ψ_t be a flow that connects $\mathbf{R}_1$ to $\mathbf{R}_0$, computing $\mathbf{u}_t(\mathbf{R}_t|\mathbf{R}_0,\mathbf{R}_1)$ simplifies to determining $\mathbf{R}_t$ along $\frac{\mathrm{d}}{\mathrm{d}t}\Psi_t(\mathbf{R}) = \dot{\mathbf{R}}_t$ (Chen & Lipman, 2023) and then taking its time derivative. Thus, we have

$$\mathbf{u}_t(\mathbf{R}_t|\mathbf{R}_0,\mathbf{R}_1) = \frac{\log_{\mathbf{R}_t}(\mathbf{R}_0)}{t} \tag{14}$$

Optimal transport. Optimal Transport (OT) conditions hold when the probability paths between two distributions are defined by a displacement map that linearly interpolates between them (Pooladian et al., 2023).

Tong et al. (2023) views the OT problem as finding a mapping that minimizes the 2-Wasserstein distance between two distributions p_1 and p_0 on $\mathbb{R}^3$, using the Euclidean distance $\|\mathbf{x}_0 - \mathbf{x}_1\|$ as the displacement cost:

$$W(p_0,p_1)_2^2 = \inf_{\pi\in\Pi}\int_{\mathbb{R}^3\times\mathbb{R}^3}\|\mathbf{x}_0-\mathbf{x}_1\|^2\mathrm{d}\pi(\mathbf{x}_0,\mathbf{x}_1) \tag{15}$$

where Π denotes the set of all joint probability measures on $\mathbb{R}^3\times\mathbb{R}^3$ with marginals p_1 and p_0. By setting $p(\mathbf{x}_0,\mathbf{x}_1) = \pi(\mathbf{x}_0,\mathbf{x}_1)$ and a Gaussian conditional probability path with mean $\mu_t = t\mathbf{x}_0 + (1-t)\mathbf{x}_1$, we have

$$\mathcal{L}_{\mathrm{OT}}^{\mathbf{X}}(\theta) = \mathbb{E}_\pi\left\|\mathbf{u}_t(\mathbf{X}_t|\mathbf{X}_0,\mathbf{X}_1) - v(\theta,t,\mathbf{X}_t)\right\|^2 \tag{16}$$

$$\mathbf{u}_t(\mathbf{x}_t|\mathbf{x}_0,\mathbf{x}_1) = \mathbf{x}_0 - \mathbf{x}_1 \tag{17}$$

with $p_t(\mathbf{x}_t) = \int\mathcal{N}(\mathbf{x}_t|t\mathbf{x}_0+(1-t)\mathbf{x}_1,\sigma^2)\pi(\mathbf{x}_0,\mathbf{x}_1)\mathrm{d}\mathbf{x}_0\mathrm{d}\mathbf{x}_1$.

Inspired by this, Bose et al. (2023) extended Equation (13) and Equation (14) to $\mathrm{SO}(3)$ using Riemannian optimal transport, with $\bar{\pi}$ being the projection of π on $\mathrm{SO}(3)$:

$$\mathcal{L}_{\mathrm{OT}}^{\mathbf{R}}(\theta) = \mathbb{E}_{\bar{\pi}}\left\|\frac{\log_{\mathbf{R}_t}(\mathbf{R}_0)}{t} - v(\theta,t,\mathbf{R}_t)\right\|^2 \tag{18}$$

B.5. $\mathrm{SE}(3)$ Invariance

$\mathrm{SE}(3)$ invariance can be achieved by consistently positioning the model at the origin (Yim et al., 2023; Bose et al., 2023; Rudolph et al., 2020).

In the context of SM, to ensure translation invariance on $\mathbb{R}^3$, one applies a projection matrix $\mathbf{P}\in\mathbb{R}^{3N\times 3N}$ that removes the center of mass $\frac{1}{N}\sum_{i=1}^N\mathbf{x}_i$. It results in an invariant measure on $\mathrm{SE}(3)^N$, denoted as $\mathrm{SE}(3)_0^N$. Since the Brownian motion on $\mathrm{SO}(3)$ and the score $\nabla_{\mathbf{R}}\log p_{T-t}$ are both rotation-invariant, Equation (4) is $\mathrm{SO}(3)$-invariant. Consequently, Yim et al. (2023) derive the following $\mathrm{SE}(3)$-invariant forward process

$$\mathrm{d}\mathbf{T}_t = \left[0, -\frac{1}{2}\mathbf{P}\mathbf{X}_t\right]\mathrm{d}t + \left[\mathrm{d}\mathbf{B}_t^{\mathrm{SO}(3)^N}, \mathbf{P}\mathrm{d}\mathbf{B}_t^{\mathbb{R}^{3N}}\right] \tag{19}$$

and its corresponding time-reverse process

$$\mathrm{d}\overleftarrow{\mathbf{R}}_t = \nabla_{\mathbf{R}}\log p_{T-t}(\overleftarrow{\mathbf{T}}_t)\mathrm{d}t + \mathrm{d}\mathbf{B}_t^{\mathrm{SO}(3)^N} \tag{20}$$

$$\mathrm{d}\overleftarrow{\mathbf{X}}_t = \mathbf{P}\left\{\frac{\overleftarrow{\mathbf{X}}_t}{2} + \nabla_{\mathbf{x}}\log p_{T-t}(\overleftarrow{\mathbf{T}}_t)\right\}\mathrm{d}t + \mathrm{d}\mathbf{P}\mathbf{B}_t^{\mathbb{R}^{3N}} \tag{21}$$

The same approach can be applied to FM. After centering and decoupling the flow on $\mathrm{SE}(3)_0^N$, a separate $\mathrm{SE}(3)$-invariant flow can be constructed for each residue in backbone[5], in which each $\mathrm{SE}(3)$-invariant measure is decomposed into a measure that is proportional to the Lebesgue measure on $\mathbb{R}^3$ (Pollard, 2001) and an $\mathrm{SO}(3)$-invariant measure (Bose et al., 2023).

[5]As the product group of N copies of $\mathrm{SE}(3)$, $\mathrm{SE}(3)_0^N$ has a geometric structure that allows global geometric operations (such as geodesic distance, exponential maps, and logarithmic maps) to be decomposed into operations on each of the N $\mathrm{SE}(3)$ groups.

B.6. Additional Losses

To prevent unrealistic fine-grained features such as steric clashes or chain breaks when learning the torsion angle ψ, Yim et al. (2023) proposed adding two additional loss functions. The first is the mean squared error (MSE) on backbone atom positions:

$$\mathcal{L}_{\text{bb}} = \frac{1}{4N} \sum_{n=1}^{N} \sum_{a \in A} \|a_n - \hat{a}_n\|^2 \tag{22}$$

where $A = \{\text{N, C, C}\alpha\text{, O}\}$. a_n and $\hat{a}_n$ are the true and predicted coordinates of backbone atom a at residue n.

The second loss penalizes deviations in local pairwise atomic distances:

$$\mathcal{L}_{\text{2D}} = \frac{\sum_{n=1}^{N} \sum_{m=1}^{N} \sum_{a,b \in A} \mathbb{1}\{d_{ab}^{nm} < 6\text{Å}\} \|d_{ab}^{nm} - \hat{d}_{ab}^{nm}\|^2}{\left(\sum_{n=1}^{N} \sum_{m=1}^{N} \sum_{a,b \in A} \mathbb{1}\{d_{ab}^{nm} < 6\text{Å}\}\right) - N} \tag{23}$$

where $d_{ab}^{nm} = \|a_n - b_m\|$ and $\hat{d}_{ab}^{nm}$ are the true and predicted distances between atoms a and b in residues n and m, respectively. The indicator function $\mathbb{1}\{d_{ab}^{nm} < 6\text{Å}\}$ limits the loss to atom pairs within 6 Å.

The complete training loss is given by

$$\mathcal{L} = \mathcal{L}^{\mathbf{R}}(\theta) + \mathcal{L}^{\mathbf{X}}(\theta) + \mathbb{1}\{t < T/4\} (\mathcal{L}_{\text{bb}} + \mathcal{L}_{\text{2D}}) \tag{24}$$

where $T = 1$ in the case of FM.

B.7. Model Architecture

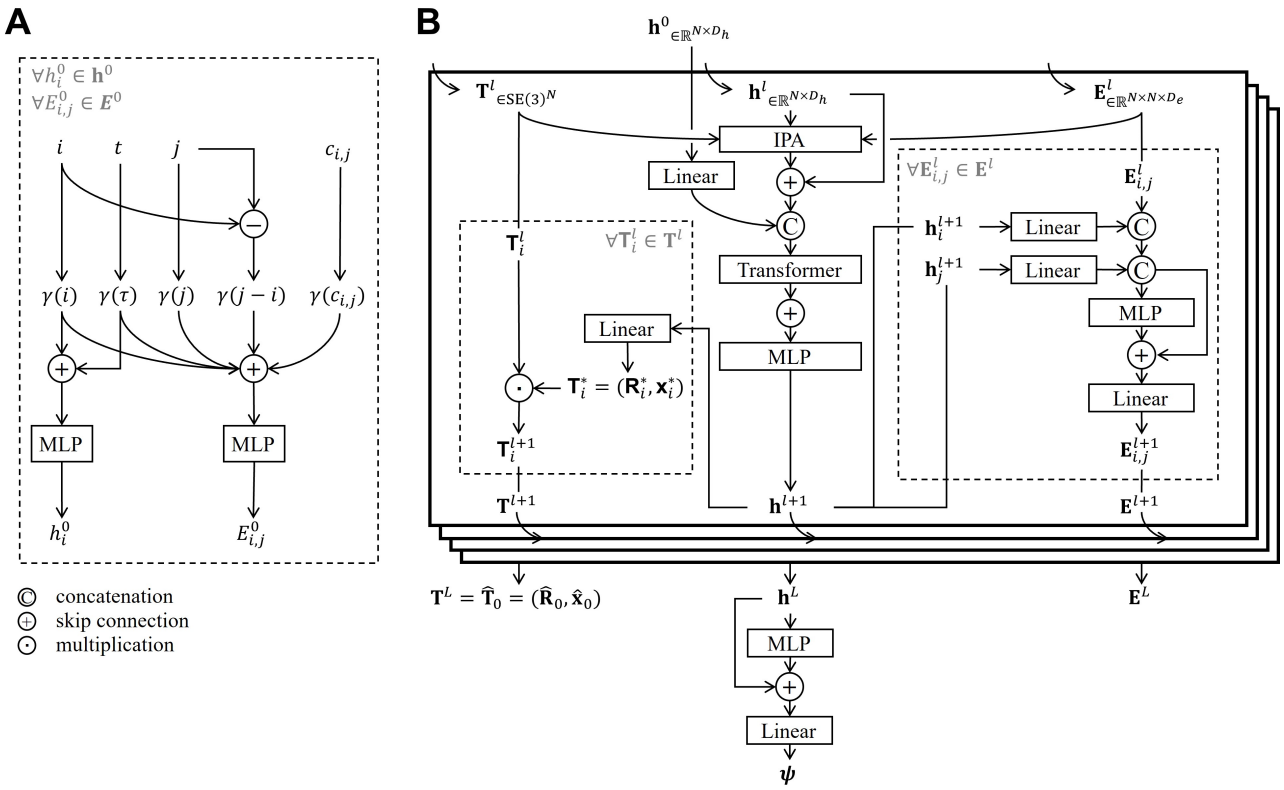

Figure 3. Overview of the (**A**) embedding module and (**B**) multi-layer network architecture.

The networks $s(\theta, t, \cdot)$ involved in SM and $v(\theta, t, \cdot)$ involved in FM models, as reviewed in Appendix B, can share a common high-level architecture (Yim et al., 2023; Bose et al., 2023; Anand & Achim, 2022).

Embeddings. Given node embedding dimensions D_h and edge embedding dimensions D_e, node embeddings $\mathbf{h} \in \mathbb{R}^{N \times D_h}$ are derived from residue indices $\mathbf{i} = \{1, \ldots, N\}$ and time-step information $\mathbf{t} = \{0, \Delta t, \ldots, T\}$, while edge embeddings $\mathbf{E} \in \mathbb{R}^{N \times N \times D_e}$ integrate additional features, such as relative sequence distances $j - i$ for any $i, j \in [1, N]$ (Figure 3A). Self-conditioning on the predicted C_α displacements is also applied:

$$c_{i,j} = \sum_{b=1}^{B} \mathbb{1}\left\{|\mathbf{x}_i^* - \mathbf{x}_j^*| < v_b\right\} \tag{25}$$

where $\mathbf{x}^*$ denotes the coordinates for C_α predicted through self-conditioning, and $v_1, \ldots, v_B$ are bins spaced uniformly from 0 to B angstroms. These initial features are encoded using multilayer perceptrons (MLPs) along with sinusoidal embeddings (Vaswani et al., 2017).

Multi-layer network. Figure 3B shows the architecture of the multi-layer neural network ($L = 4$ layers used in our experiments). At each layer l, the network takes node embeddings $\mathbf{h}^l$, edge embeddings $\mathbf{E}^l$, and rigid transformations $\mathbf{T}^l$ as input, applying the Invariant Point Attention (IPA) introduced by Jumper et al. (2021) to enable spatial attention. Transformer from Vaswani et al. (2017) models interactions along the chain structure. The network's update procedure remains invariant under $\mathrm{SE}(3)$ transformations due to the inherent $\mathrm{SE}(3)$-invariance of the IPA.

The output $\mathbf{T}^L$ from the final layer serves as the predicted frame, denoted as $\hat{\mathbf{T}}_0 = (\hat{\mathbf{R}}_0, \hat{\mathbf{x}}_0)$. Consequently, for SM, we have the following scores predictions based on Equation (8) and Equation (9):

$$\forall s(\theta, t, \mathbf{R}_t) \in s(\theta, t, \mathbf{R}_t), \quad s(\theta, t, \mathbf{R}_t) = \nabla_{\mathbf{R}} \log p_{t|0}(\mathbf{R}_t | \hat{\mathbf{R}}_0) \tag{26}$$

$$= \frac{\mathbf{R}_t}{\omega(\hat{\mathbf{R}}_0^\top \mathbf{R}_t)} \log\{\hat{\mathbf{R}}_0^\top \mathbf{R}_t\} \frac{\partial_\omega f(\omega(\mathbf{R}_0^\top \mathbf{R}_t), t)}{f(\omega(\mathbf{R}_0^\top \mathbf{R}_t), t)} \tag{27}$$

$$\forall s(\theta, t, \mathbf{x}_t) \in s(\theta, t, \mathbf{X}_t), \quad s(\theta, t, \mathbf{x}_t) = \nabla_{\mathbf{x}} \log p_{t|0}(\mathbf{x}_t | \hat{\mathbf{x}}_0) \tag{28}$$

$$= \frac{e^{-t/2}\hat{\mathbf{x}}_0 - \mathbf{x}_t}{1 - e^{-t}} \tag{29}$$

From Equation (11) and Equation (14), we have the following for FM with OT:

$$\forall v(\theta, t, \mathbf{R}_t) \in v(\theta, t, \mathbf{R}_t), \quad v(\theta, t, \mathbf{R}_t) = \mathbf{u}_t(\mathbf{R}_t | \hat{\mathbf{R}}_0, \mathbf{R}_1) \tag{30}$$

$$= \frac{\log_{\mathbf{R}_t}(\hat{\mathbf{R}}_0)}{t} \tag{31}$$

$$\forall v(\theta, t, \mathbf{x}_t) \in v(\theta, t, \mathbf{X}_t), \quad v(\theta, t, \mathbf{x}_t) = \mathbf{u}_t(\mathbf{x}_t | \hat{\mathbf{x}}_0) \tag{32}$$

$$= \frac{\hat{\mathbf{x}}_0 - (1 - \sigma)\mathbf{x}_t}{1 - (1 - \sigma)t} \tag{33}$$

Torsion angle $\hat{\boldsymbol{\psi}} = \{\hat{\psi}_1, \ldots, \hat{\psi}_N\} = \boldsymbol{\psi}/\|\boldsymbol{\psi}\| \in \mathrm{SO}(2)^N$ is predicted with $\mathbf{h}^L$ and $\mathbf{E}^L$.

C. Protein Families Involved in This Study

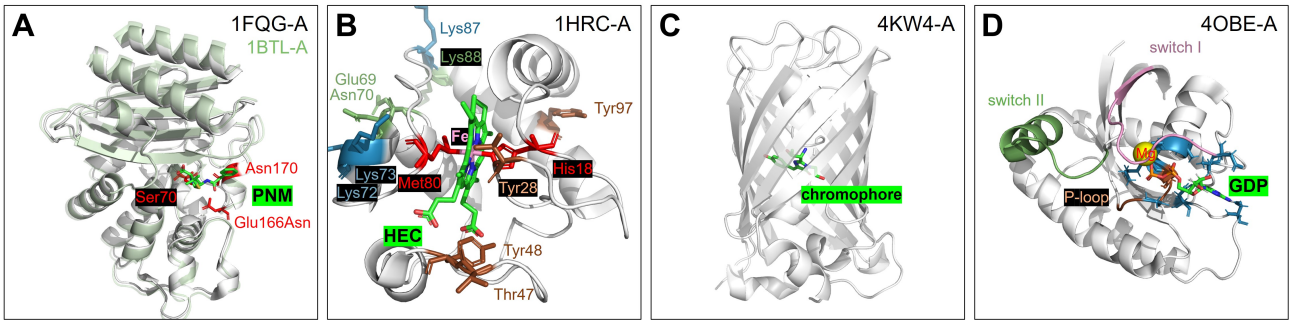

Figure 4. (**A**) Overlay of WT *E. coli* TEM1 (PDB: 1BTL; Jelsch et al. (1993); green) and its E166N acylated intermediate (PDB: 1FQG; Brown et al. (2009); white) with penicillin (PNM). (**B**) WT *E. caballus* heart cytochrome *c* (PDB: 1HRC; (Dickerson et al., 1967)) with heme C (HEC). (**C**) *A. victoria* GFP and its chromophore (PDB: 4KW4; Barnard et al. (2014)). (**D**) GDP-bound *H. sapiens* KRas protein (PDB: 4OBE; (Hunter et al., 2014)).

C.1. β-lactamases

β-lactamases are enzymes that deactivate β-lactam antibiotics by hydrolyzing their β-lactam ring, contributing significantly to bacterial resistance (Lee et al., 2016). Inhibiting these enzymes can restore antibiotic efficacy (Behzadi et al., 2020). Rapid diversification driven by the evolution of bacterial resistance makes β-lactamases ideal targets for protein modeling studies. For this study, we gathered structural data on 1,578 unique monomeric β-lactamases and their variants across Ambler classes (A, B, C, and D) from the BLDB (Naas et al., 2017) and the Protein Data Bank (PDB) (Berman, 2000).

Class A β-lactamases are the most prevalent, with conserved active-site residues Ser^{70}, Glu^{166}, and Asn^{170} coordinating the hydrolytic water for deacylation (Tooke et al., 2019; Brown et al., 2009). Figure 4A shows the WT β-lactamase TEM1 (1BTL) alongside its acylated E166N intermediate (1FQG).

C.2. Cytochrome *c*

Cytochrome *c* is a water-soluble protein ($\sim$ 12 kDa) essential for ATP synthesis in mitochondria and intrinsic apoptosis (Kashyap et al., 2021; Ow et al., 2008). It also serves as an independent marker for apoptosis in several cancers (Li et al., 2001; Way et al., 2004). Despite variations across species, its core structure and function are conserved. We obtained structural data for 498 unique cytochrome *c* proteins and variants from the PDB.

Figure 4B shows horse cytochrome *c*, where a hydrophobic shell surrounds the heme group, with only $\sim 7.5\%$ of the surface available for electron transfer (Bushnell et al., 1990). The hydrophobic environment and iron coordination by His^{18} and Met^{80} maintain a high redox potential (260 mV) (Salemme, 1977). Phosphorylation occurs at Thr^{28}, Thr^{47}, Tyr^{48}, and Tyr^{97} (Hüttemann et al., 2011), while Lys^{72}, Lys^{73}, and Lys^{87} bind phospholipids (Kagan et al., 2009). The ATP-binding pocket involves Glu^{69}, Asn^{70}, Lys^{88}, and Lys^{72}, Lys^{86}, Lys^{87} (McIntosh et al., 1996).

C.3. Green fluorescent proteins (GFP)

GFP, first isolated from *Aequorea victoria*, fluoresces green when stimulated by specific wavelengths. Its core structure is an 11-strand β-barrel enclosing the chromophore (Figure 4C) (Remington, 2011). Various mutants have been engineered to enhance or modify its properties, including enhanced GFP (Cormack et al., 1996), superfolder GFP (Pédelacq et al., 2005), and color variants like YFP (Ormö et al., 1996) and BFP (Glaser et al., 2016). We collected structural data for 448 GFPs and variants from the PDB to explore this diversity.

C.4. Ras Proteins

Ras proteins, a subgroup of the small GTPase superfamily, act as molecular switches, cycling between GTP-bound (active) and GDP-bound (inactive) states to regulate cell proliferation, differentiation, migration, and apoptosis (Ladygina et al., 2011; Simanshu et al., 2017; Weinmann & Ottow, 2007). They play a key role in signaling from the cell surface to downstream pathways. Mutations that keep Ras proteins in an active state drive excessive cell growth and malignancy, making Ras

inhibition a promising cancer treatment strategy (Singh & Lingham, 2002). We focused on the most common cancer-related Ras proteins (HRas, KRas, and NRas) (Cox, 2002) and obtained 511 experimentally derived structures from the PDB.

In human KRas (Figure 4D), the switch I and II regions form the key interface for effector and regulator binding (Pantsar, 2020). These regions are highly flexible, with conformations depending on GTP or GDP binding. Cancer-related mutations frequently occur in the P-loop and switch II (Pantsar, 2020).

C.5. Sequence Length Distributions in Experimentally Derived and Generated Protein Structures

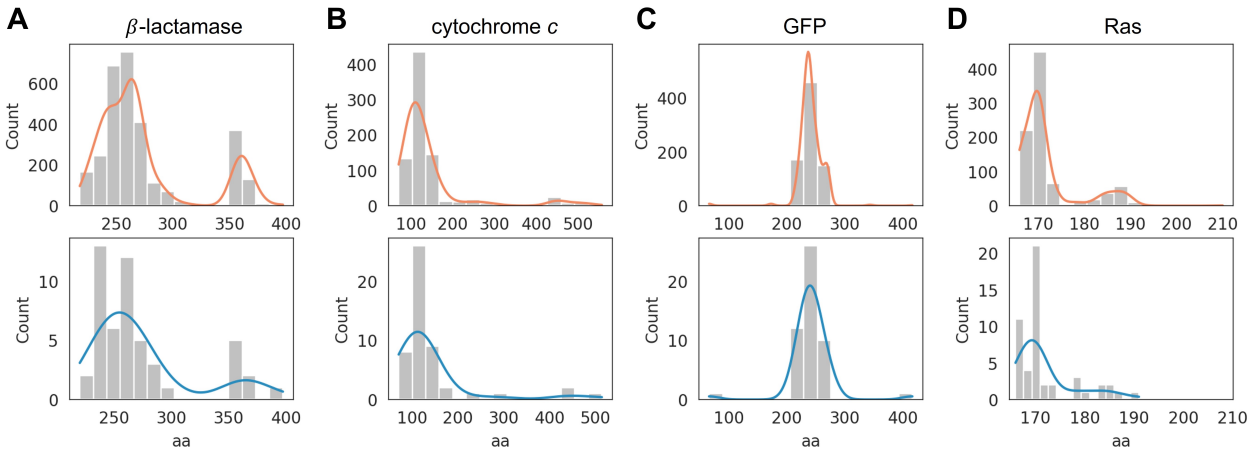

Figure 5. Distributions of amino acid sequence lengths (aa) for the experimentally derived protein structures used for training (top row; orange) and the 50 backbone structures generated by each model (bottom row; blue). Target sequence lengths for the generated structures were sampled from the training data distribution, shown here for (**A**) β-lactamase, (**B**) cytochrome c, (**C**) GFP, and (**D**) Ras.

D. Dihedral Angles ψ and ϕ Distributions

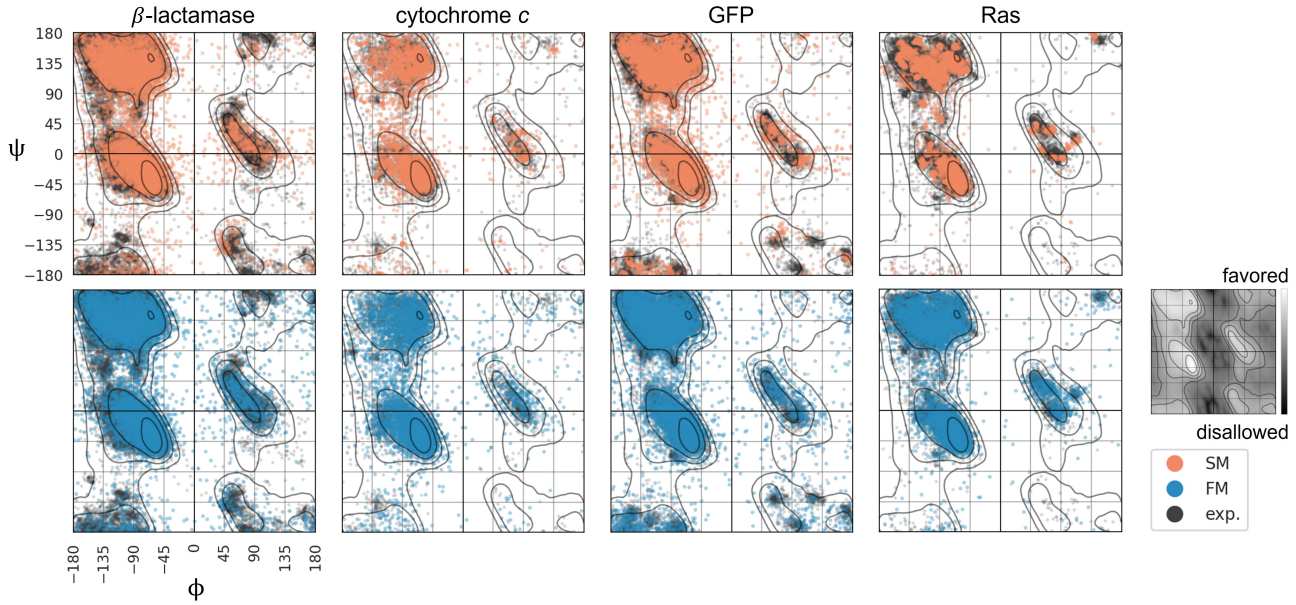

Figure 6. Ramachandran plots comparing dihedral angles ψ and ϕ distributions for generated versus experimentally derived proteins; inset shows favored (light) and disallowed (dark) regions.

E. Conserved Residue Consistency

E.1. Evolutionary Rates Mapped onto Structures

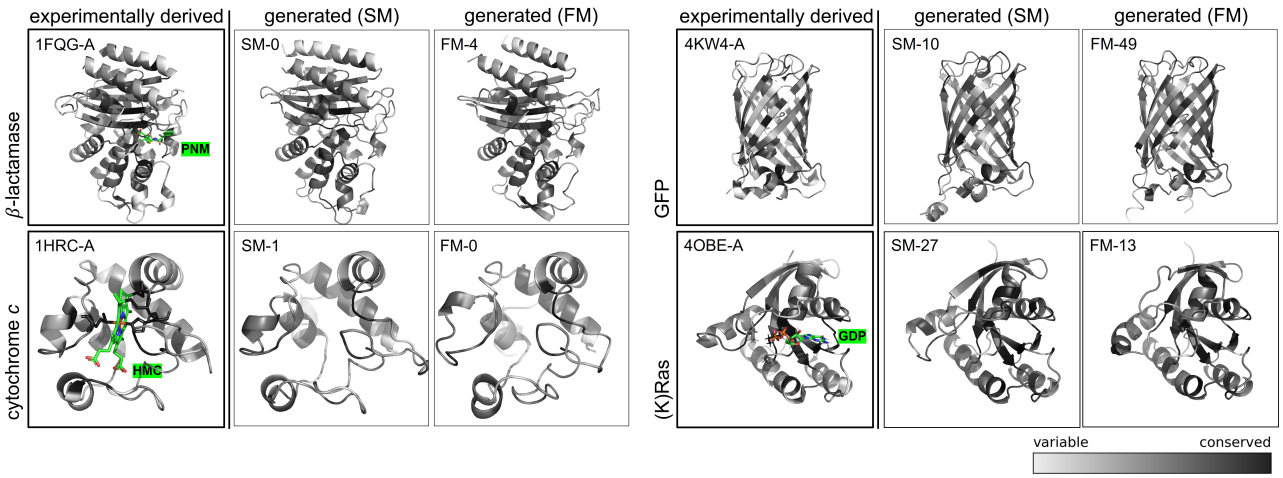

Figure 7. Normalized evolutionary rates mapped onto structures, with white for rapidly evolving positions and black for conserved ones. Experimentally derived structures are highlighted with bold outlines. Refer to Figure 4 and Figure 12 for the relevant key residues and the locations of the ligand-binding pockets (represented by PNM, HMC and GDP).

E.2. Evolutionary Rates Mapped onto Sequences

Figure 8 shows that, except for β-lactamase, the FM-generated sequences show a slightly higher average pairwise distance than those of the SM, indicating greater diversity. One possible explanation is that the experimentally derived β-lactamase structures used for training (from four distinct Ambler classes) have more variability than the other three protein families (as also reflected in Figure 9) and that SM is more sensitive to structural variability.

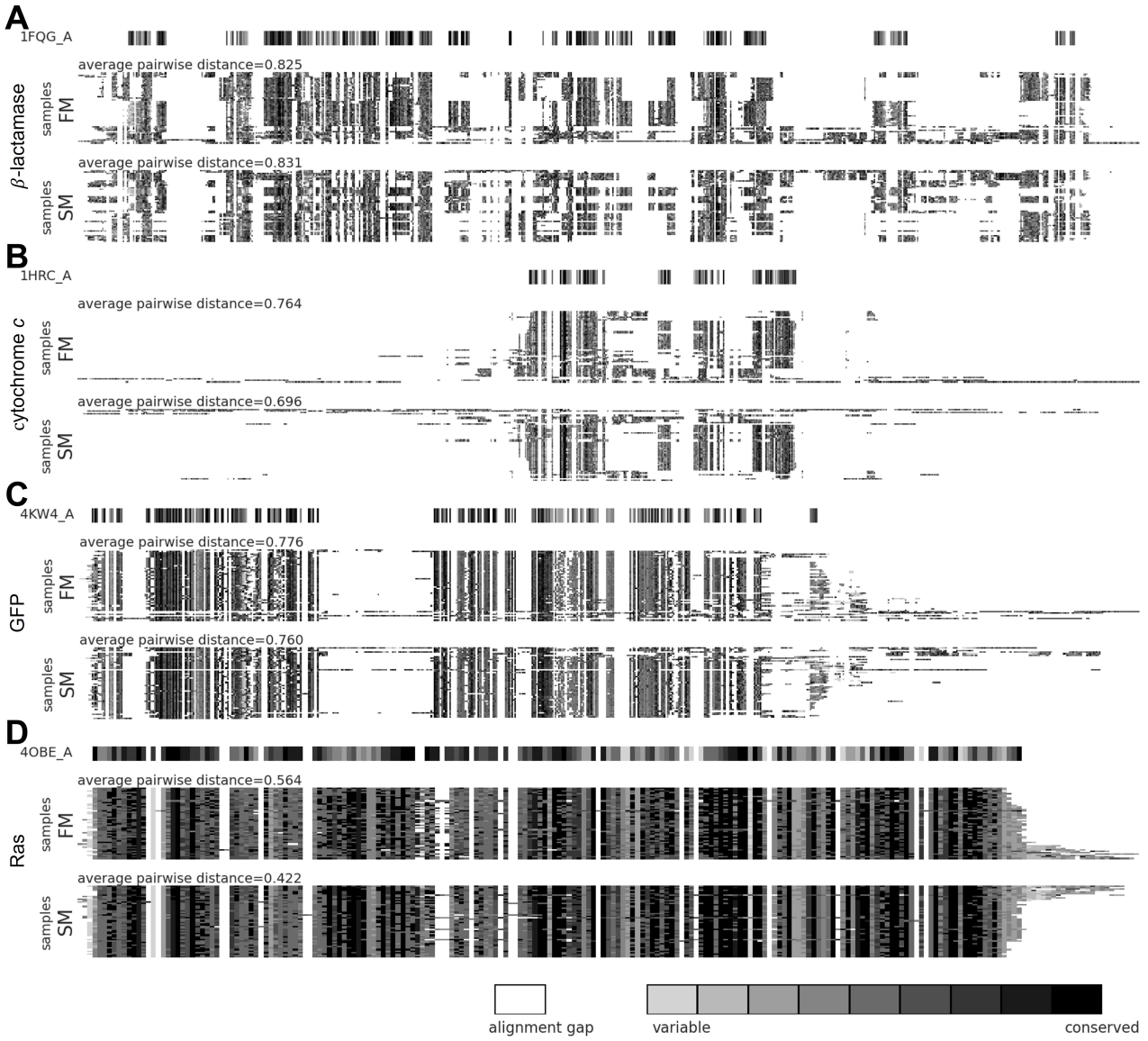

Figure 8. Multiple sequence alignment of generated samples and experimentally derived reference proteins (1FQG, 1HRC, 4XWA, 4O8E) using Clustal Omega, with normalized evolutionary rates overlaid. Rapidly evolving positions are highlighted in light gray, conserved regions in black, and alignment gaps in white. In addition, a comparative analysis of variability is presented: average pairwise distances, derived from the Clustal Omega distance matrix, denote greater variability when larger values are observed.

F. Summary Structural Phylogenetic Tree

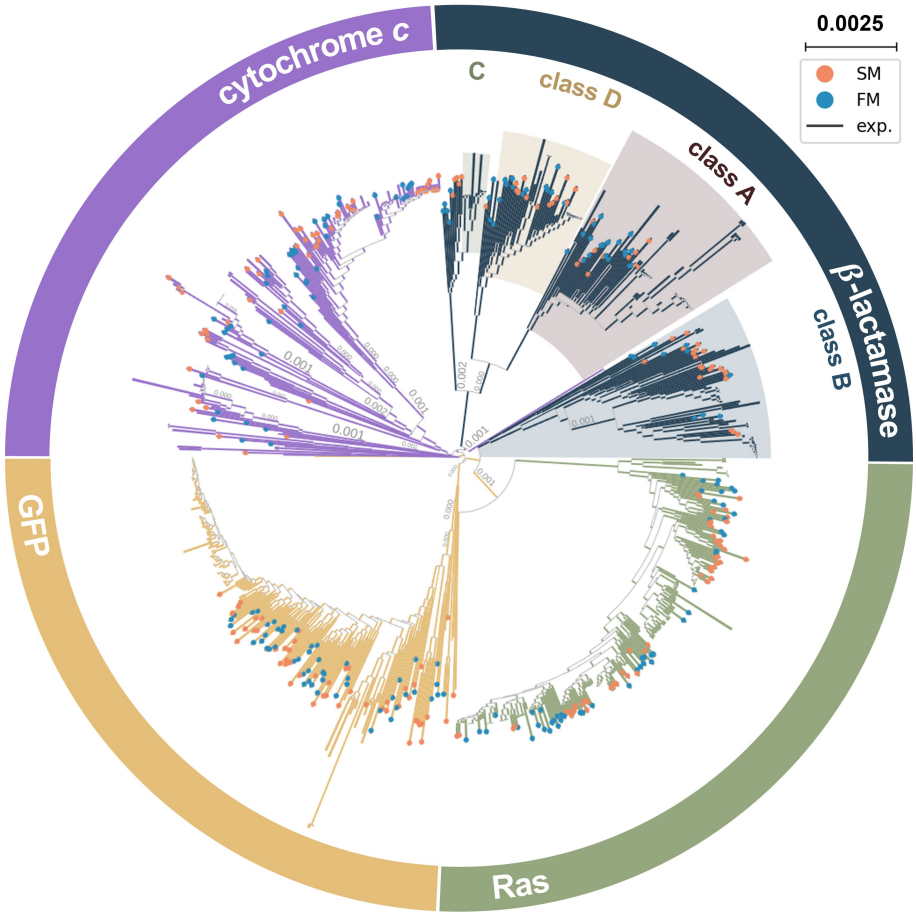

Figure 9. Summary structural phylogenetic tree constructed using the Q_{score} and the 3Di alphabet. Branch colors distinguish families, with orange and blue nodes representing SM- and FM-generated structures, respectively. In β-lactamases, distinct Ambler classes are differentiated by unique background colors.

G. Structure-informed Trees versus Sequence-based Phylogenetic Trees

Retrieving known taxonomic lineages. Following Moi et al. (2023), we retrieved taxonomic lineages for each experimentally derived sequence and structure within every protein family using the UniProt API (Patient et al., 2008), assuming that most genes evolve in a manner that mirrors the species tree with only occasional instances of gene loss or duplication.

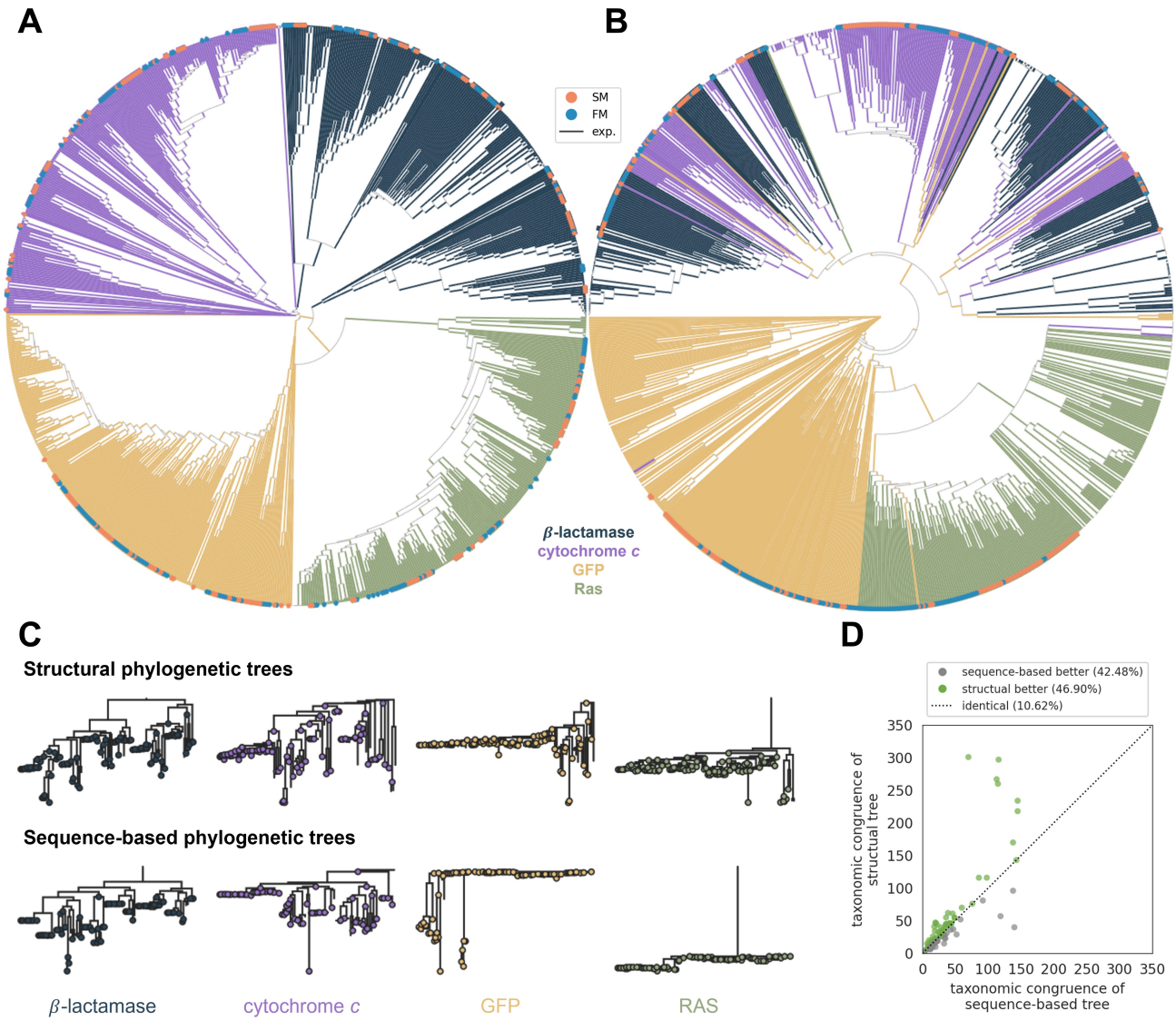

Figure 10. (**A**) Ultrametric summary structural phylogenetic tree constructed using the Q_{score} and the 3Di alphabet. (**B**) Ultrametric sequence-based phylogenetic tree constructed using the Clustal Omega and the FastTree pipeline. Different protein families are differentiated by distinct colored branches. Nodes in orange and blue represent structures generated by SM and FM, respectively. (**C**) Phylogenetic trees of distinct protein families. Top: summary structural phylogenetic tree constructed using the Q_{score} metric and the 3Di alphabet. Bottom: sequence-based phylogenetic tree inferred using the Clustal Omega and FastTree pipeline. (**D**) Taxonomic congruence score for each node in the sequence and structure trees. On average, structural trees exhibit higher taxonomic congruence than sequence-based trees.

Using Q_{score} in structural phylogenetics. For any two structures with N_1 and N_2 residues, Q_{score} is computed with TM-align (Zhang, 2005) as:

$$Q_{\text{score}} = \frac{N_{\text{align}}^2}{N_1 N_2} \times \frac{1}{1 + \left(\frac{\text{RMSD}}{R_0}\right)^2} \tag{34}$$

where N_{align} is the number of aligned residues, RMSD is the root-mean-square deviation of atomic positions, and R_0 (set to $4\mathring{\text{A}}$) balances the contributions of RMSD and N_{align}.

Taxonomic congruence score (TCS). Tan et al. (2015) proposed the use of the TCS to assess how well a phylogenetic tree's topology agrees with the known taxonomy, arguing that TCS is an unbiased measure of tree quality. Typically, trees with higher average TCS (structure trees in Figure 10D) are considered to have more accurate topologies. In this study, we evaluated the congruence between the given trees and the established taxonomy lineages derived by UniProt. Subsequently, we provide a brief overview of the bottom-up TCS implementation as described by Moi et al. (2023):

For any node x in the tree, $s(x)$ is the set of taxonomic lineage labels present in the subtree rooted at x. If x is a leaf, then $s(x)$ is defined to be the set of lineage labels (its taxonomic classification) for that leaf. If x is an internal node with children $\{y_1, \cdots, y_N\}$, then $s(x) = \bigcap_{i=1}^{N} s(i)$.

For each node x, the function $C(x)$ quantifies the congruence of that node's grouping with taxonomy: $C(x) = |s(x)| + |s(p)|$, where p is the parent nodes of x, and $|\cdot|$ denotes the size of a set.

The overall taxonomic congruence score for the entire tree is obtained by summing the contributions of all leaves. To compare congruence across trees of different sizes, the raw total score is typically normalized.

H. Evaluation of Side-Chain Homology Modeling

We used homology modeling to add side-chains to the generated protein backbones, evaluating them using PROCHECK (Laskowski et al., 1993; 1996) and WHAT_CHECK (Hooft et al., 1996) to correct or exclude those not meeting the expectations. Specifically:

Planarity. Planar side-chains, such as those in phenylalanine, tyrosine, tryptophan, and histidine, are essential for stability and function. Conformations lacking expected planarity were discarded.

Asparagine, glutamine, histidine flips. Asparagine, glutamine, and histidine side-chains can experience terminal flips, altering key interactions. WHAT_CHECK was used to evaluate and, if necessary, adjust side-chain orientations to have more stable interactions.

Torsion angles. Side-chain torsion angles (χ angles) were assessed, focusing on χ_1 (rotation around C_α to the first side-chain atom) and χ_2 (rotation to the second side-chain atom) to prevent spatial clashes. Conformations in uncommon χ-angle regions were excluded.

Bond lengths and angles. Unusual bond lengths and angles may indicate strain and modeling errors, potentially disrupting interactions. Conformations with such issues were discarded.

Other parameters. Side-chains with abnormal torsion angles, atypical aromatic bonding angles, or unusual proline puckering were also discarded.

I. Molecular Dynamics and Blind Docking Simulations

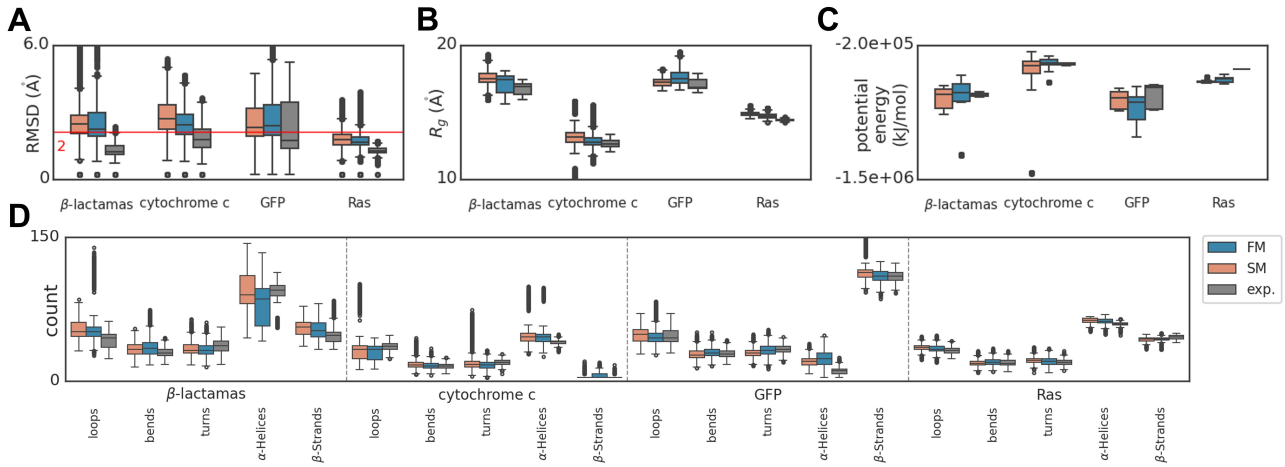

Figure 11. Stability assessment of MD simulations across proteins using various metrics. Distributions of (**A**) RMSD, (**B**) radius of gyration (R_g), (**C**) potential energy, and (**D**) secondary structure counts throughout the simulation. Interquartile ranges and whiskers show metric variation; high-quality structures have medians close to experimentally derived values with narrow ranges.

For experimentally derived structures, crystallographic water and unnecessary small molecules were removed. For generated structures, missing side-chains were added via homology modeling (Section 2.5). After adding hydrogen atoms using Reduce2 (Grosse-Kunstleve et al., 2002) and confirming no missing atoms, each protein was centered at the origin.

Simulations were performed with GROMACS (Abraham et al., 2015), using the all-atom CHARMM36 force field (July 2022 version) (Vanommeslaeghe et al., 2009; Vanommeslaeghe & MacKerell, 2012; Vanommeslaeghe et al., 2012; Yu et al., 2012; Soteras Gutiérrez et al., 2016).

I.1. Molecular Dynamics Setup for Stability Assessment of Generated Structures

Proteins were placed in an octahedral simulation box with a minimum distance of 1.5 nm between the protein and the box boundaries. Prior to solvation, energy minimization was performed in vacuum using the steepest descent method (max 30,000 steps, step size 0.01 nm, convergence 2 kJ/(mol·nm)) to resolve steric clashes and geometric inconsistencies. Neighbor searching used a grid-based method with a search radius of 1.2 nm.

In accordance with GROMACS 2024 documentation, we applied the following configurations in the MD parameter (.mdp) files. Van der Waals interactions were handled using a cutoff method, while long-range electrostatic interactions were calculated using the Particle Mesh Ewald (PME) method.

```
constraints      = h-bonds
cutoff-scheme    = Verlet
vdwtype          = cutoff
vdw-modifier     = force-switch
rlist            = 1.2
rvdw             = 1.2
rvdw-switch      = 1.0
coulombtype      = PME
rcoulomb         = 1.2
DispCorr         = no
```

After solvating the system with water using the TIP3P model, we added Na^+ and Cl^- ions to achieve a physiological concentration of 150 mM and to neutralize the system's total charge. Energy minimization was then conducted to resolve steric clashes and optimize the geometry, with potential energy and maximum force monitored to ensure they reached acceptable thresholds.

The next step involved equilibrating the solvent and ions around the protein. We chose the leap-frog integrator for the simulations and applied the LINCS algorithm to constrain hydrogen bonds. Equilibration involved two stages. In the first

stage, we performed a 500 ps NVT[6] equilibration (250,000 steps with a 2 fs time step). Temperature control was managed using the V-rescale thermostat, with the system divided into two groups: (1) protein and (2) water + ions, both set to a target temperature of 310 K to simulate physiological conditions. In the second stage, we carried out a 500 ps NPT [7] equilibration with pressure coupling enabled. The pressure was regulated using the C-rescale method with isotropic coupling. The target pressure was 1.0 bar, with a compressibility of 4.5×10^{-5} bar^{-1} and a pressure coupling time constant of 0.5 ps.

Following equilibration, we conducted a 10 ns production simulation (5,000,000 steps with a 2 fs time step), during which all position restraints were removed. This allowed us to observe and analyze the system's dynamic behavior over time, in order to access its stability. Full details of the MD parameter files can be found in Software and Data.

I.2. Molecular Dynamics Setup for Conformational Analysis of Protein-Ligand Complexes

The receptor and ligand were saved as separate coordinate files to prepare their respective topologies. The receptor topology was prepared as in Appendix I.1. For the ligand, hydrogen atoms were added using OpenBabel (O'Boyle et al., 2011), and topology was generated via the CGenFF server (Vanommeslaeghe et al., 2009; Vanommeslaeghe & MacKerell, 2012; Vanommeslaeghe et al., 2012). The receptor and ligand topologies, along with force-field-compatible coordinate files, were then combined to construct the complete complex system.

The MD workflow for complexes followed the same steps in Appendix I.1. Complexes were placed in an octahedral simulation box, energy-minimized in vacuum, solvated in water, and neutralized with Na^+ and Cl^- ions to 150 mM. A second energy minimization was then performed on the solvated system.

During equilibration, positional restraints were applied to the ligand to prevent unnecessary displacement in the initial stages of the simulation. Additionally, to minimize interference from temperature fluctuations of the ligand on the overall simulation, we defined two temperature coupling groups: (1) the receptor and ligand as one group, and (2) the solvent and ions as the other. Other equilibration settings followed Appendix I.1.

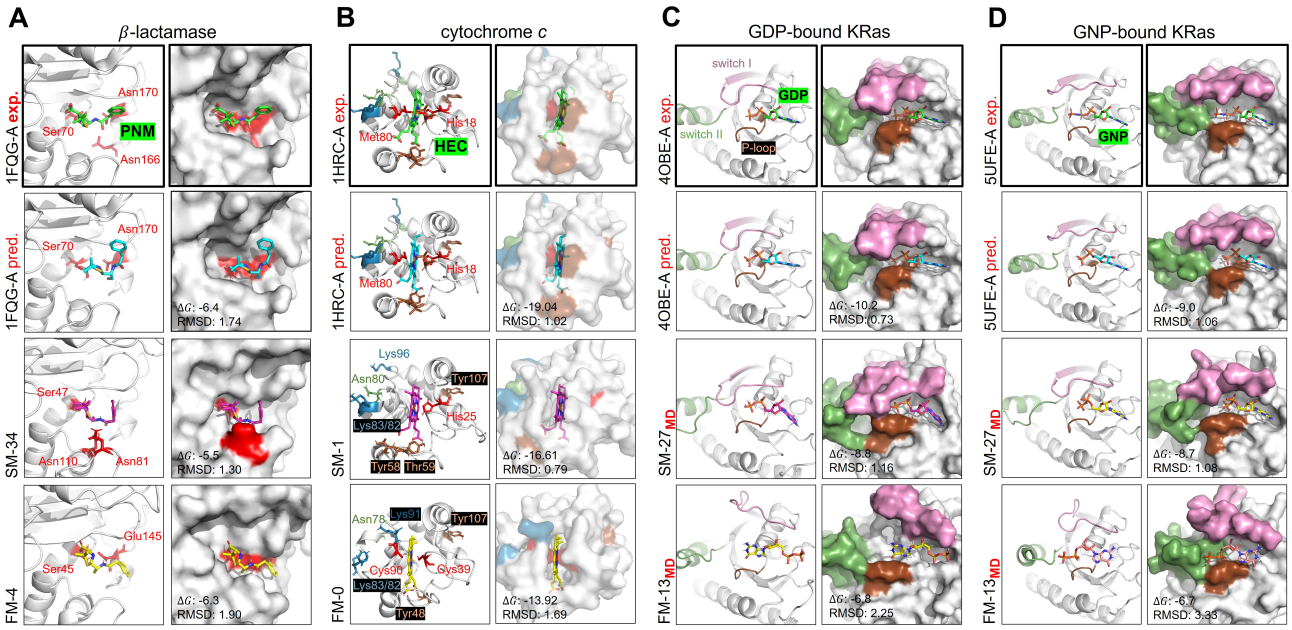

Figure 12. Comparison of experimentally derived binding modes (bold boxes; first row) with predictions from blind docking simulations for receptors binding to family-specific ligands. Ligands are colored as green (ground truth), cyan (predicted using experimentally derived receptors), magenta (predicted using SM samples), and yellow (predicted using FM samples). ΔG (kcal/mol) and RMSD (Å) quantify binding affinities and deviations from the experimental poses. "MD" labels protein-ligand complex MD simulations after docking.

After equilibration, restraints were removed, and a 10 ns production simulation was conducted to analyze the dynamic

[6]Constant number of particles, volume, and temperature.
[7]Constant number of particles, pressure, and temperature.

behavior and conformational changes in the complexes.

I.3. Protein-ligand Blind Docking

Similarly, crystallographic water and unwanted molecules were removed from experimentally derived structures, and missing side-chains were added to generated structures via homology modeling. Receptor structures were prepared using AutoDock Tools (Morris et al., 2009), with polar hydrogens added, Kollman charges assigned, and any missing atoms repaired. For receptors within the same family, we prepared a shared ligand file, adding hydrogen atoms and assigning Gasteiger charges. A large grid box, typically 80 to 110 Å per side, was defined to cover the entire protein surface.

Using these settings, we performed blind docking with AutoDock Vina (Trott & Olson, 2009; Eberhardt et al., 2021), generating up to 25 binding modes with a maximum energy difference of 5 kcal/mol and an exhaustiveness level of 20. The binding mode with the lowest binding free energy was selected as the final result.

J. Applicability and Limitations of Deep Generative Protein Design Guidelines

- It is hard to apply strict physical constraints to very flexible proteins in generative process. In addition, using only steric exclusions without considering environmental factors such as the lipid bilayer in membrane proteins or interfaces in large assemblies can make the designs less accurate (Winnifrith et al., 2024).
- If a target protein's stability or fold depends critically on bound cofactors or oligomeric assembly, designing sequences in isolation can be unreliable. According to Krishna et al. (2024), generating and screening candidate sequences without accounting for these interactions will often fail to reproduce the native structure or stability observed in the full cofactor- or multimer-associated complex.
- For truly new folds or functions, there are no known conserved residues to guide design. In families where conserved sites are spread out or have moved over evolution, keeping those sites can block creative changes. In practice, the design of novel proteins remains a low-success "attritional" problem with success only in rare, isolated cases (Greener et al., 2018).
- Standard force fields often do not model metal ions, sugar attachments, or membrane effects accurately. As a result, refinement steps may produce structures that differ from what actually happens in the lab.
- For novel protein folds with no known homologs, template-based modeling or comparison is difficult. Even state-of-the-art predictors admit substantially reduced accuracy when no homologous structure exists (Jumper et al., 2021).
- Conventional MD simulations are limited to relatively short timescales due to computational cost. For example, simulating $\sim 50,000$ atoms (a modest protein) for ~ 1 μs can take days on a GPU (Hollingsworth & Dror, 2018). Moreover, simulations of systems with membranes, metal centers, or covalent modifications often suffer from force-field artifacts or setup uncertainties.
- Docking methods that do not model backbone flexibility often cannot accommodate the large backbone and side-chain rearrangements required for binding, leading to unreliable predictions (Lexa & Carlson, 2012).

K. Generated Structures

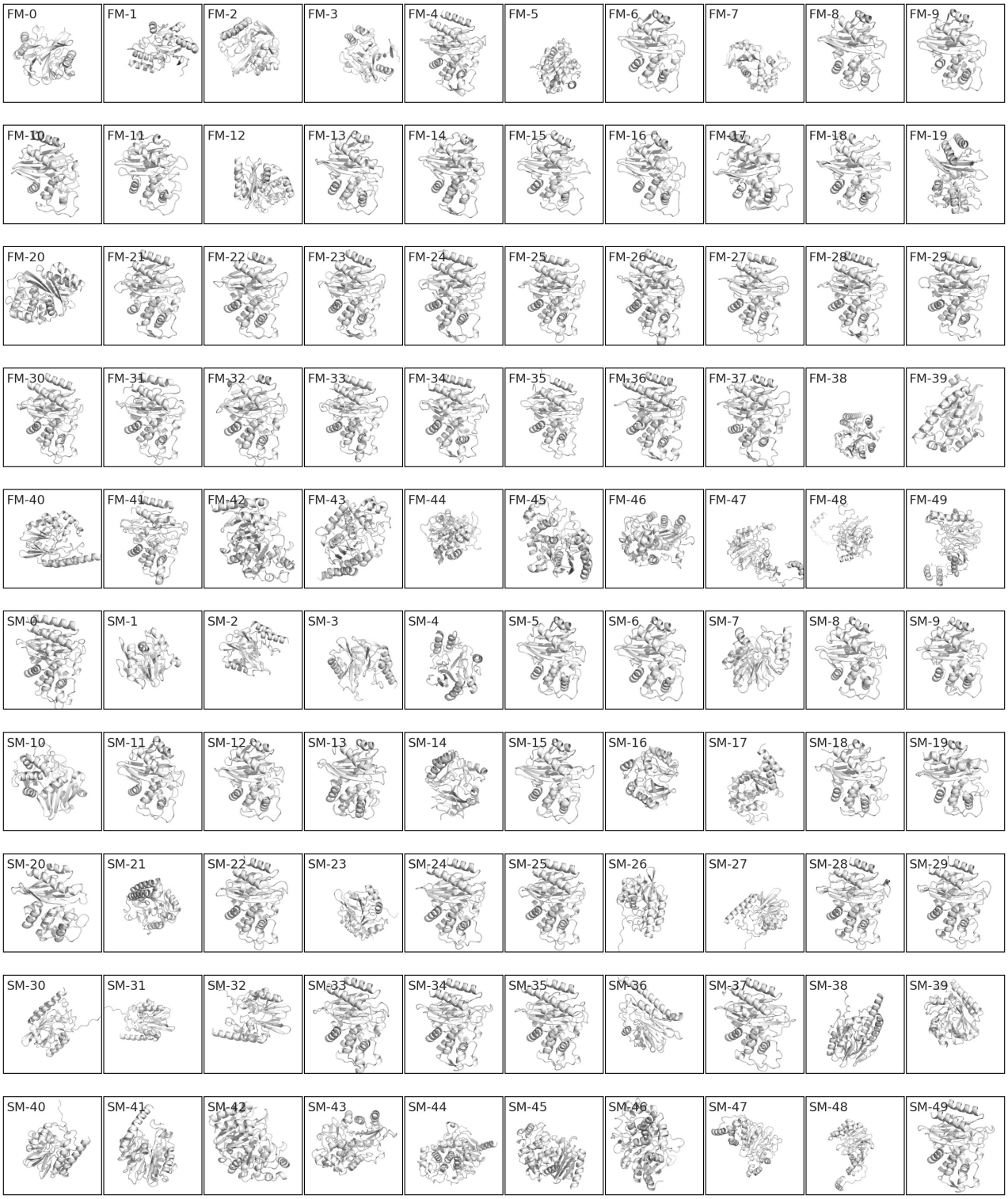

Figure 13. 50 β-lactamase-like protein backbones generated using score matching and 50 using flow matching.

Figure 14. 50 cytochrome *c*-like protein backbones generated using score matching and 50 using flow matching.

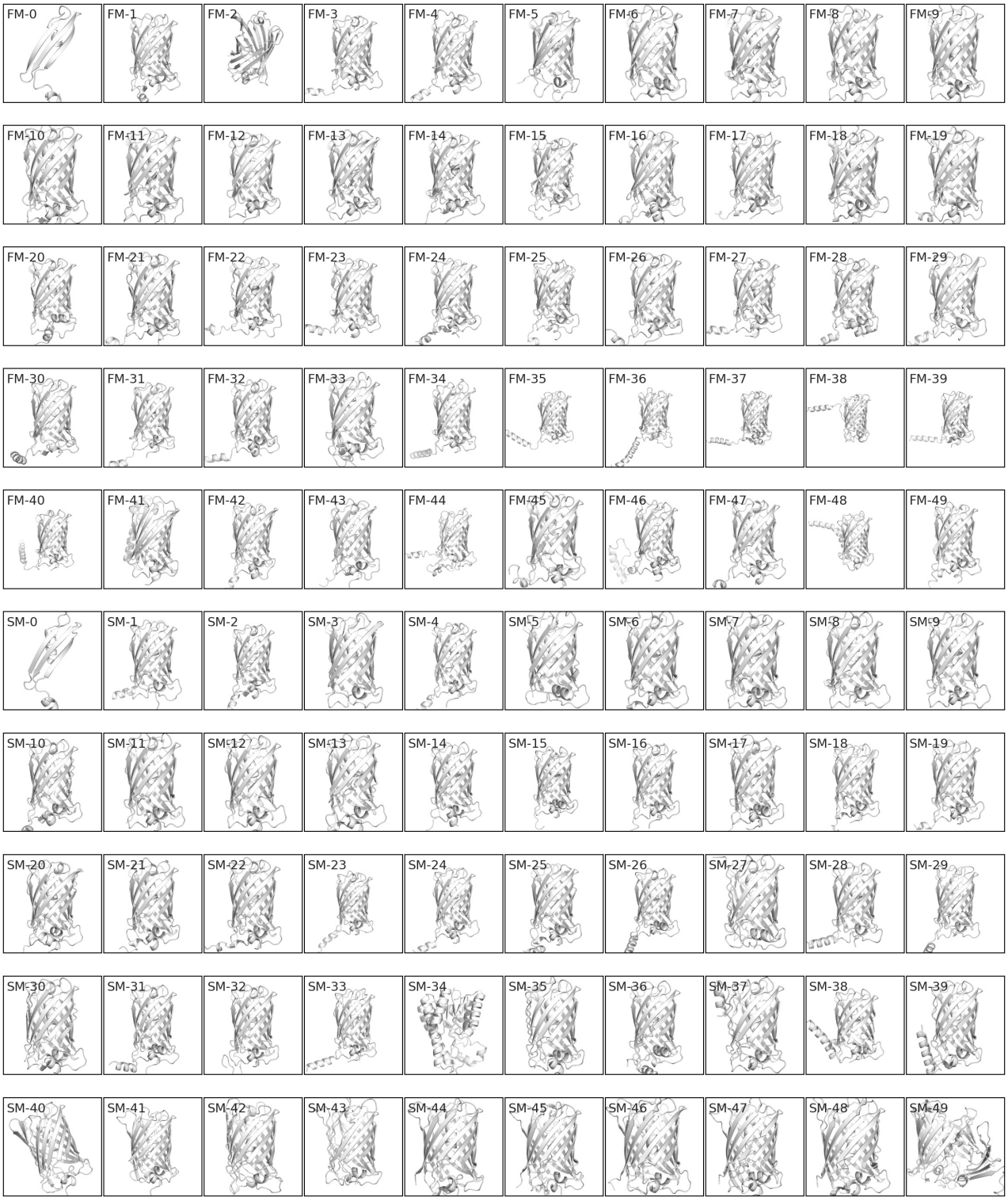

Figure 15. 50 GDP-like protein backbones generated using score matching and 50 using flow matching.

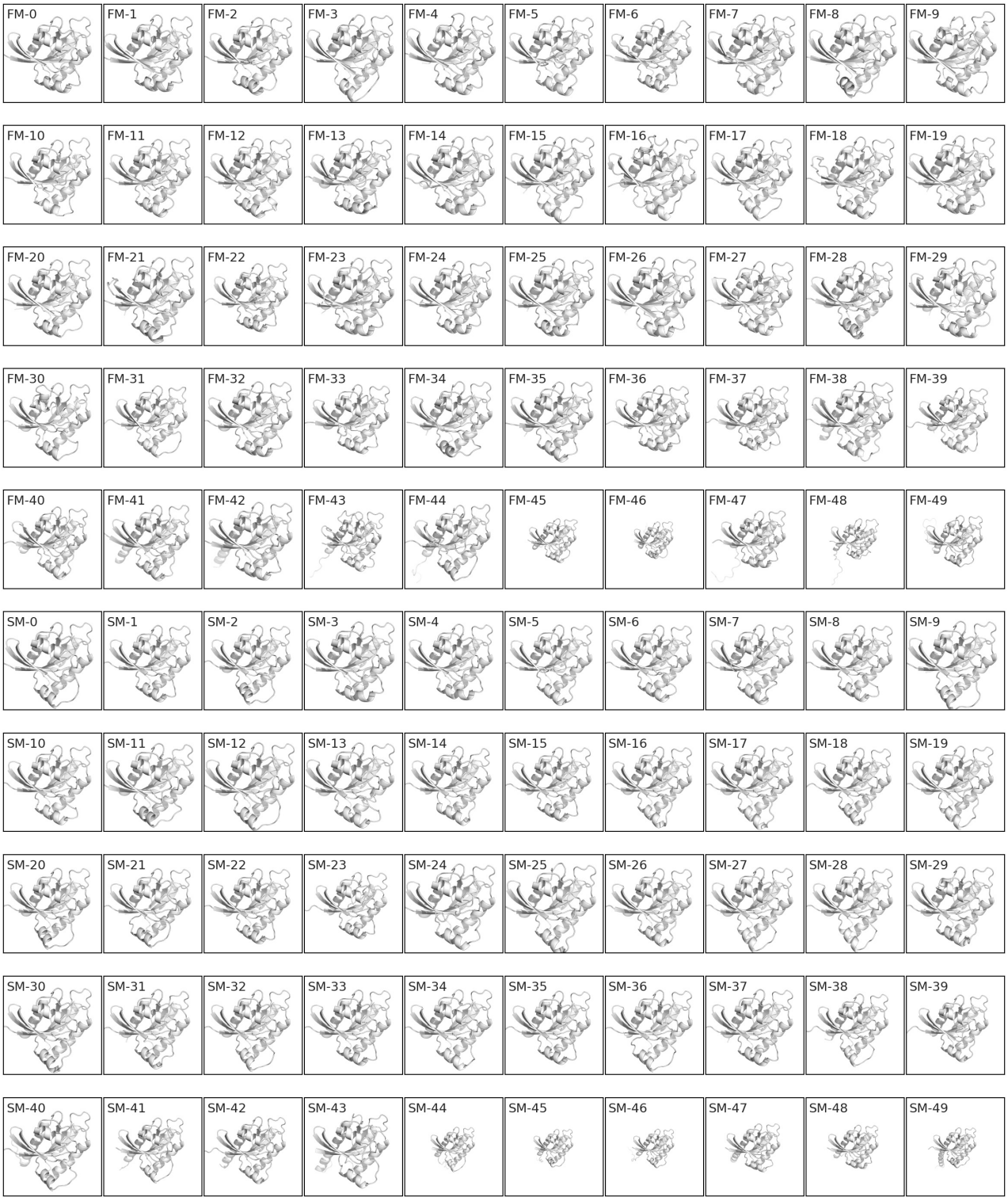

Figure 16. 50 Ras-like protein backbones generated using score matching and 50 using flow matching.