Integrating Protein Language Models and All-Atom Generative Design for Predicting Mutation-Induced Stability Changes

Milena Boucsein
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
milena.work@buffalo.edu

Mutia Marpaung
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
mutiamarpaung539@uc.edu

Polly L. Commins
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
polly.commins61@missouri.edu

Abstract

Predicting mutation-induced stability changes is a foundational challenge for protein engineering, therapeutic development, biomanufacturing, and synthetic biology. Recent advances in protein language models have made it possible to infer evolutionary and functional constraints directly from sequence data, while all-atom generative design has begun to model molecular geometry, side-chain packing, and physicochemical interactions with increasing fidelity. This paper examines how these two research programs can be integrated into a system-level architecture for predicting mutation-induced stability changes. Rather than treating sequence-based prediction and structure-based generation as competing methods, the paper argues that they are complementary layers of a broader socio-technical infrastructure that must address data curation, model interoperability, uncertainty communication, deployment constraints, sustainability, robustness, and governance. The analysis emphasizes architectural trade-offs between broad sequence pretraining and high-resolution all-atom modeling, between zero-shot generalization and experimental calibration, and between centralized foundation-model services and distributed domain-specific fine-tuning. It also considers the policy implications of systems capable of proposing or evaluating mutations that could alter protein stability, function, or host range. The paper concludes that progress will depend not only on improved predictive accuracy but also on auditable data pipelines, reproducible evaluation regimes, equitable access, environmental accountability, and responsible translational pathways. A robust integrated system must therefore be understood as an institutional and infrastructural achievement as much as a computational one.

Keywords

protein language models; all-atom generative design; mutation-induced stability changes; protein engineering; AI governance; robust deployment; structural bioinformatics; socio-technical systems.

1. Introduction

The prediction of mutation-induced stability changes occupies a central position in contemporary computational biology because it connects molecular structure, evolutionary history, laboratory validation, and translational application. A single amino acid substitution can alter folding energetics, oligomeric assembly, binding affinity, immunogenicity, and industrial robustness. For decades, experimental methods such as directed evolution and biophysical characterization have provided the gold standard for assessing such effects, but their cost and throughput limit the scale at which variants can be explored. Computational prediction has therefore become an essential complement, enabling researchers to prioritize mutations before committing laboratory resources. The emergence of highly accurate protein structure prediction demonstrated that deep learning could capture complex spatial relationships from evolutionary and structural data [1]. Subsequent work on protein language models revealed that large-scale sequence pretraining can encode structural, functional, and evolutionary signals without explicit supervision [2], [3]. These developments created a new foundation for mutation-effect prediction, but they also exposed persistent gaps between sequence-level inference and atom-level mechanistic explanation.

At the same time, all-atom generative design has advanced rapidly. Generative models now attempt to produce protein backbones, sequences, side-chain conformations, and molecular assemblies under geometric and physicochemical constraints [8], [9], [10]. Such methods promise to move beyond scoring existing variants toward proposing new variants or even new folds with desired stability profiles. However, all-atom generation remains computationally intensive, sensitive to data quality, and difficult to evaluate in realistic deployment settings. Protein language models, by contrast, are comparatively efficient and can generalize across families, but they often lack explicit treatment of atomic detail and may fail when stability depends on subtle packing, metal coordination, or solvent-mediated interactions. The integration of these approaches is therefore not merely a model-combination problem. It is a systems problem involving data infrastructure, model architecture, validation pipelines, uncertainty quantification, governance, and sustainability.

This paper argues that integrating protein language models and all-atom generative design for mutation-induced stability prediction requires a layered architecture. The sequence layer supplies evolutionary priors and broad generalization. The structural layer supplies geometric and energetic constraints. The generative layer proposes and refines variants. The evaluation layer calibrates predictions against experimental and biophysical evidence. The governance layer manages access, safety, bias, and environmental impact. Each layer introduces trade-offs that cannot be resolved by isolated benchmarks. For example, a model that performs well on average may be unreliable for rare protein families, membrane proteins, or mutations that disrupt metal-binding sites. A system optimized for cloud-scale inference may exclude academic laboratories with limited computational resources. A generative model that proposes highly stable variants may also produce sequences with unknown immunogenicity or ecological risk. Consequently, the design of integrated stability prediction systems must be assessed through both technical and socio-technical criteria.

The paper proceeds by first reviewing the conceptual foundations of protein language models and all-atom generative design. It then develops a system architecture for integrated mutation-induced stability prediction, followed by an analysis of data infrastructure and governance. Subsequent sections examine deployment, robustness, sustainability, fairness, policy, evaluation, and translational pathways. The conclusion synthesizes these themes and identifies a research agenda for building systems that are accurate, accountable, and equitable.

2. Conceptual Foundations of Protein Language Models and All-Atom Generative Design

Protein language models treat amino acid sequences as a language and learn statistical dependencies across residues through large-scale pretraining. Models trained on hundreds of millions of sequences can capture conservation patterns, co-evolutionary couplings, and implicit structural constraints [2], [3]. These representations have been used for fitness prediction, variant effect estimation, and evolutionary forecasting [4], [5]. In the context of mutation-induced stability changes, protein language models provide a powerful prior: mutations that violate deeply conserved constraints are more likely to be destabilizing, while substitutions observed frequently across homologs may be tolerated or even stabilizing. Zero-shot mutation-effect prediction demonstrated that language models can rank variants without task-specific supervision, although their performance varies by protein family and mutation type [6]. The strength of this approach lies in breadth and scalability, but its weakness lies in interpretability and mechanistic specificity. A language model may assign a low probability to a mutation without revealing whether the effect arises from hydrophobic core disruption, electrostatic repulsion, or backbone strain.

All-atom generative design addresses this limitation by explicitly representing atoms, bonds, side-chain rotamers, and nonbonded interactions. Early graph-based generative models for protein design demonstrated that deep learning could propose sequences conditioned on structure [8]. Diffusion and denoising models subsequently extended this capability to backbone generation and sequence-structure co-design [9], [10]. These methods can generate novel protein structures and sequences with desired geometric properties, and they can be fine-tuned for tasks such as motif scaffolding, binder design, and stability optimization. In principle, all-atom generative design can evaluate mutation-induced stability changes by generating alternative side-chain packing arrangements and estimating energetic consequences. In practice, however, the computational cost is high, the search space is enormous, and the accuracy of energy evaluation depends on force fields, solvent models, and conformational sampling. Moreover, generative models are often trained on structures that are biased toward soluble, well-folded, and crystallizable proteins, which limits their applicability to membrane proteins, intrinsically disordered regions, and engineered sequences.

The complementarity between protein language models and all-atom generative design is therefore substantial but not automatic. Language models can propose mutations that are evolutionarily plausible and globally compatible with a protein family. All-atom generative models can test whether those mutations fit the local geometry and physicochemical environment. Conversely, all-atom models can generate candidate structures or side-chain conformations that are then scored by language models for evolutionary plausibility. This bidirectional coupling is more than an ensemble. It creates a system in which sequence-level priors constrain structural search, and structural realism constrains sequence generation. The integration challenge is to design interfaces between these layers that preserve uncertainty, allow error propagation, and remain computationally tractable.

A further conceptual issue is the distinction between stability, function, and fitness. Mutation-induced stability changes are often correlated with function and fitness, but they are not identical. A mutation can stabilize a protein while reducing catalytic activity, altering allosteric regulation, or changing binding specificity. Protein language models trained on evolutionary data may conflate stability with fitness because evolution selects for function in a specific environment. All-atom generative models may predict improved packing while

neglecting dynamic motions required for activity. An integrated system must therefore represent multiple objectives and avoid collapsing them into a single score. This requires architectural support for multi-task learning, uncertainty-aware ranking, and explicit trade-off analysis.

3. System Architecture for Integrated Stability Prediction

A system-level architecture for integrated mutation-induced stability prediction can be understood as a set of interacting layers rather than a single model. The sequence representation layer ingests wild-type and mutant sequences, produces embeddings, and estimates evolutionary plausibility. The structural representation layer encodes available experimental structures, predicted structures, or generated conformations. The all-atom generative layer proposes side-chain packing, backbone adjustments, and alternative conformations under geometric constraints. The scoring and calibration layer combines sequence-level and structure-level evidence with biophysical heuristics and experimental data. The workflow orchestration layer manages compute, caching, provenance, and reproducibility. The governance layer controls access, monitors misuse, and maintains audit trails. Each layer has distinct failure modes and trade-offs.

At the center of this architecture is the need for interoperability between model families. Protein language models often output per-residue or per-mutation embeddings, while all-atom generative models operate on coordinate tensors and atom types. Bridging these representations requires learned adapters or shared latent spaces. Recent work on all-atom multimodal Bayesian flow attempts to unify sequence, structure, and geometric constraints within a single generative process, suggesting that future systems may not treat sequence and structure as separate modules [13]. However, unified models also introduce complexity: they can be harder to interpret, more expensive to train, and more difficult to update when new data become available. A modular architecture may be easier to govern and maintain, but it risks error accumulation across interfaces. The design choice between unification and modularity is therefore a central architectural trade-off.

Another trade-off concerns generalization versus specialization. A general protein language model can be applied to many families, but it may underperform on specific targets such as antibodies, enzymes, or membrane transporters. Fine-tuning on domain-specific data can improve accuracy but may reduce robustness to distribution shift. All-atom generative models face a similar tension: broad pretraining on the Protein Data Bank enables diverse generation, but specialized fine-tuning on stable scaffolds can produce better local packing at the cost of novelty. An integrated system could support multiple specialized heads or adapters, with a routing mechanism that selects the appropriate model based on protein class, mutation type, and available structural data. Routing itself introduces governance questions: if certain models are more accurate but less accessible, the system may reproduce existing inequities in computational biology.

Uncertainty quantification is another architectural requirement. Mutation-induced stability changes are often small relative to experimental error, and predictions can be sensitive to input structure quality. An integrated system should distinguish between aleatoric uncertainty arising from experimental noise and epistemic uncertainty arising from limited data or model mismatch. It should also communicate when a prediction is based on a predicted structure rather than an experimental structure, because the latter carries different confidence. Ensemble methods, Bayesian approximations, and conformal prediction can provide calibrated uncertainty, but they increase computational cost and complexity. In deployment,

uncertainty estimates must be presented in ways that support decision-making without implying false precision.

The architecture must also support iterative refinement. A typical use case begins with a wild-type structure, a set of candidate mutations, and a desired stability objective. The system generates initial predictions, ranks variants, and may propose additional mutations. All-atom generative design can then refine side-chain packing and estimate local strain. Protein language models can rescore the refined variants for evolutionary plausibility. This loop can continue until convergence or resource limits. The orchestration layer must manage state, version models, track data provenance, and ensure that the same input produces reproducible outputs. Without such infrastructure, integrated systems become difficult to validate and trust.

4. Data Infrastructure, Curation, and Governance

Data infrastructure is the substrate of integrated stability prediction. Protein language models depend on large sequence databases, while all-atom generative models depend on structural databases, experimental stability measurements, and simulation trajectories. These data sources differ in quality, coverage, licensing, and bias. Sequence databases are vast but noisy, with redundant entries, annotation errors, and uneven taxonomic representation. Structural databases are high quality but biased toward proteins that are amenable to crystallography, cryo-electron microscopy, or nuclear magnetic resonance. Stability databases are smaller and often heterogeneous, with different experimental conditions, assays, and definitions of stability. Integrating these sources requires careful curation, normalization, and provenance tracking.

Curation is not a purely technical task. It involves decisions about which organisms, protein families, and mutation types are represented. If a stability prediction system is trained primarily on soluble bacterial proteins, it may perform poorly on human membrane proteins or therapeutic antibodies. If it relies on structures determined at low temperature, it may misestimate flexibility and entropy. If it excludes negative results or failed designs, it may overestimate the success rate of generative proposals. Governance of data infrastructure must therefore include transparency about inclusion criteria, metadata standards, and known limitations. It must also address licensing and access, because some databases restrict commercial use or require attribution.

The governance of integrated systems extends beyond data to model access and dual use. A system that predicts mutation-induced stability changes could be used to stabilize therapeutic proteins, improve industrial enzymes, or enhance the environmental persistence of a pathogen. The same capabilities that support beneficial protein engineering can lower barriers to harmful modification. Policy responses must therefore be proportionate and not stifle legitimate research. Options include tiered access, know-your-customer requirements for cloud services, screening of synthesis orders, and international coordination on biological design tools. The challenge is to design governance mechanisms that are technically informed, adaptable to rapid model improvement, and equitable across regions and institutions.

Provenance and auditability are essential for both scientific and governance purposes. An integrated prediction should be traceable to the input sequence, structural model, training data version, model checkpoint, and calibration dataset. If a prediction later proves inaccurate, researchers should be able to diagnose whether the error arose from data, architecture, or deployment. Audit trails also support regulatory review in clinical and environmental contexts. However, comprehensive provenance can conflict with privacy, proprietary data, and

computational efficiency. A balance must be struck through standardized metadata, cryptographic hashing, and federated learning approaches that keep sensitive data local while sharing model updates.

5. Deployment, Robustness, and Sustainability

Deployment of integrated stability prediction systems must account for heterogeneous computational environments. Large language models and all-atom generative models often require graphics processing units, high memory, and substantial storage. Cloud deployment can provide scalability, but it raises costs, data transfer issues, and dependence on commercial providers. On-premises deployment may offer greater control and privacy, but it can be unaffordable for smaller laboratories. A sustainable deployment strategy should therefore support multiple modalities, including distilled models for local inference, application programming interfaces for large-scale generation, and hybrid workflows that combine local preprocessing with remote refinement.

Robustness is closely tied to deployment. A model that performs well on a benchmark may fail when inputs are noisy, structures are incomplete, or mutations are outside the training distribution. Robustness testing should include adversarial mutations, rare protein families, missing residues, alternate conformations, and environmental conditions. It should also assess sensitivity to input structure quality. Because predicted structures are increasingly used as inputs, systems must be robust to errors in those predictions. This requires uncertainty-aware pipelines that can downweight unreliable structural regions and flag predictions that depend on low-confidence geometry.

Sustainability has both computational and environmental dimensions. Training and deploying large protein models consume energy and contribute to carbon emissions. All-atom generative design can be especially expensive because it may require many sampling steps per candidate. Sustainable system design should consider model compression, efficient sampling, caching, and scheduling. It should also consider the full life cycle, including data storage, hardware manufacturing, and electronic waste. In academic and industrial settings, sustainability metrics should be transparent so that researchers can compare methods not only by accuracy but also by energy and cost. Otherwise, the field may optimize for benchmark performance while externalizing environmental costs.

Maintenance is another deployment concern. Protein databases grow, new structures are solved, and experimental stability data accumulate. Models must be updated without invalidating prior results or breaking reproducibility. This requires versioning, regression testing, and clear documentation of changes. In regulated contexts, model updates may require revalidation. A sustainable system must therefore treat maintenance as a first-class design requirement rather than an afterthought. It must also plan for the eventual retirement of models and the migration of data to durable formats.

6. Fairness, Policy, and Socio-Technical Implications

Fairness in integrated stability prediction is often overlooked because the field frames itself as objective and technical. However, access to data, compute, and expertise shapes who can use these systems and who benefits from them. Well-resourced institutions can fine-tune large models, generate millions of candidates, and validate them experimentally. Smaller laboratories may be limited to public checkpoints and limited compute. This asymmetry can widen gaps in therapeutic development, agricultural biotechnology, and industrial enzyme engineering. Fairness also concerns representation: if models are trained disproportionately on

proteins from a few model organisms, their predictions may be less reliable for neglected diseases, extremophiles, or indigenous knowledge systems.

Policy implications extend to intellectual property, biosafety, and environmental release. If generative systems propose novel stable proteins, questions arise about patentability, ownership of training data, and liability for harmful outcomes. Biosafety frameworks must adapt to the pace of model development. Some proposals emphasize DNA synthesis screening, while others focus on publication norms, export controls, or international treaties. An integrated system for mutation-induced stability prediction may not itself be dangerous, but it can contribute to a broader ecosystem of biological design tools. Governance should therefore be ecosystem-wide rather than narrowly focused on a single model [20].

The socio-technical dimension also includes communication with non-expert stakeholders. Predictions of stability changes are probabilistic and context-dependent, but they may be presented as definitive scores in user interfaces. This can lead to overconfidence among researchers, investors, and policymakers. Transparent documentation, uncertainty displays, and limitations sections are necessary but insufficient. Institutions must cultivate norms of responsible interpretation and ensure that validation remains a collaborative process between computational and experimental scientists. In clinical translation, regulatory agencies will require evidence beyond retrospective benchmarks, including prospective validation and manufacturing reproducibility.

Finally, the policy environment must support both innovation and precaution. Overly restrictive rules could push research underground or concentrate capabilities in a few jurisdictions. Under-regulation could enable misuse or environmental harm. Adaptive governance, involving scientists, ethicists, security experts, and affected communities, offers a path forward. It should be informed by technical realities, such as the difficulty of distinguishing beneficial and harmful mutations, and by social values, such as equity, transparency, and accountability.

7. Evaluation, Validation, and Translational Pathways

Evaluation of integrated stability prediction systems must go beyond average error on curated benchmarks. A comprehensive evaluation framework should include prospective validation on unseen proteins, cross-family generalization, robustness to structural uncertainty, calibration of confidence estimates, and comparison with experimental measurements under standardized conditions. It should also assess computational cost, energy use, and reproducibility. Benchmarks that overrepresent small, soluble, monomeric proteins can create misleading impressions of progress. New benchmarks should include membrane proteins, multimeric assemblies, disulfide-rich peptides, and proteins with cofactors or post-translational modifications.

Validation is inherently interdisciplinary. Computational predictions must be tested through biophysical assays, directed evolution, and structural characterization. All-atom generative models can propose mutations, but experimentalists must determine whether the resulting proteins are stable, soluble, and functional. Protein language models can prioritize variants, but their rankings must be calibrated against measured stability changes. This requires close collaboration between computational and experimental laboratories, as well as data-sharing agreements that allow results to inform future models. Negative results are particularly valuable and should be reported and curated.

Translational pathways vary by application. In therapeutic protein engineering, stability prediction may support developability assessment, formulation, and manufacturing. In industrial enzymology, it may guide thermostability and solvent tolerance. In agriculture, it may influence crop resilience or food processing. Each domain has distinct regulatory, economic, and ethical contexts. A system that performs well in one domain may not transfer directly to another. Translation therefore requires domain-specific fine-tuning, validation, and governance. It also requires engagement with end users, including clinicians, manufacturers, and farmers, to ensure that the system addresses real needs rather than benchmark abstractions.

8. Future Directions and Research Agenda

Future research should focus on unified representations that preserve interpretability while capturing all-atom detail. Bayesian flow and multimodal generative models offer promising directions, but they must be evaluated for robustness, scalability, and governance readiness [13]. Another priority is uncertainty-aware decision support. Systems should not only predict stability changes but also explain why a mutation is likely to be destabilizing and what evidence would reduce uncertainty. This requires integration of mechanistic biophysics with data-driven inference.

Federated and privacy-preserving learning could expand access to proprietary or sensitive data without centralizing control. Distillation and quantization could reduce the computational burden of large models, making them more accessible. Sustainable artificial intelligence practices should be embedded in model development, including energy reporting and carbon-aware scheduling. Governance research should develop adaptive frameworks that can keep pace with model capabilities, including mechanisms for international coordination and equitable benefit sharing.

Finally, the field should invest in education and workforce development. Integrated stability prediction requires expertise in protein biophysics, machine learning, high-performance computing, data governance, and science policy. Training programs that bridge these domains will be essential for building systems that are both technically excellent and socially responsible. The next generation of tools should be designed not only to predict but also to inform, audit, and empower a broad community of researchers and stakeholders.

9. Conclusion

Integrating protein language models and all-atom generative design for predicting mutation-induced stability changes is a promising but complex undertaking. Protein language models provide evolutionary priors and broad generalization, while all-atom generative design provides geometric and physicochemical detail. Together, they can support more accurate and mechanistically grounded predictions than either approach alone. However, integration is not simply a matter of connecting two models. It requires a layered architecture, interoperable representations, uncertainty quantification, curated data infrastructure, robust deployment, sustainable computation, and adaptive governance. The system must be evaluated not only by predictive accuracy but also by reproducibility, accessibility, fairness, environmental impact, and policy alignment. As the field advances, the most successful systems will be those that treat technical and socio-technical dimensions as inseparable. They will combine state-of-the-art modeling with transparent data practices, responsible access controls, and collaborative validation. In doing so, they can help realize the promise of computational protein engineering while mitigating the risks that accompany increasingly capable biological design tools.

References

1. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589. https://doi.org/10.1038/s41586-021-03819-2

2. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., ... Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15), e2016239118. https://doi.org/10.1073/pnas.2016239118

3. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., ... Rost, B. (2022). ProtTrans: Toward understanding the language of life through self-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 7112-7127. https://doi.org/10.1109/TPAMI.2021.3095381

4. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. Nature Methods, 16(12), 1315-1322. https://doi.org/10.1038/s41592-019-0598-1

5. Hie, B. L., Yang, K. K., & Kim, P. S. (2021). Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. Cell Systems, 12(3), 243-253. https://doi.org/10.1016/j.cels.2021.02.001

6. Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., & Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. Advances in Neural Information Processing Systems, 34, 29287-29303.

7. Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., & Rives, A. (2021). Transformer protein language models are unsupervised structure learners. International Conference on Learning Representations.

8. Ingraham, J., Garg, V. K., Barzilay, R., & Jaakkola, T. (2019). Generative models for graph-based protein design. Advances in Neural Information Processing Systems, 32, 15820-15831.

9. Anand, N., Achim, T., & Huang, P. S. (2022). Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. arXiv preprint arXiv:2205.15019. https://doi.org/10.48550/arXiv.2205.15019

10. Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., ... Baker, D. (2023). De novo design of protein structure and function with RFdiffusion. Nature, 620(7976), 1089-1100. https://doi.org/10.1038/s41586-023-06415-8

11. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., ... Baker, D. (2022). Robust deep learning-based protein sequence design using ProteinMPNN. Science, 378(6615), 49-56. https://doi.org/10.1126/science.add2187

12. Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L., & Dror, R. (2020). Learning from protein structure with geometric vector perceptrons. International Conference on Learning Representations.

13. Wu, H., Song, Y., Zhang, Z., Zhang, Z., Zhou, H., Ma, W. Y., & Liu, J. (2026). Rationalized All-Atom Protein Design with Unified Multi-Modal Bayesian Flow. Advances in Neural Information Processing Systems, 38, 3019-3050.

14. Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., ... Rives, A. (2022). Learning inverse folding from millions of predicted structures. International Conference on Machine Learning, 162, 8946-8970.

15. Notin, P., Dias, M., Frazer, J., Hurtado, D. M., Gomez, A. N., Marks, D., & Gal, Y. (2023). Tranception: Protein fitness prediction with autoregressive transformers and inference-time retrieval. International Conference on Machine Learning, 162, 16990-17017.

16. Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., ... Marks, D. S. (2021). Disease variant prediction with deep generative models of evolutionary data. Nature, 599(7883), 91-95. https://doi.org/10.1038/s41586-021-04043-8

17. Riesselman, A. J., Ingraham, J. B., & Marks, D. S. (2018). Deep generative models of genetic variation capture the effects of mutations. Nature Methods, 15(10), 816-822. https://doi.org/10.1038/s41592-018-0138-4

18. Kuhlman, B., & Bradley, P. (2019). Advances in protein structure prediction and design. Nature Reviews Molecular Cell Biology, 20(11), 681-697. https://doi.org/10.1038/s41580-019-0163-x

19. Trippe, B. L., Yim, J., Tischer, D., Baker, D., Broderick, T., Barzilay, R., & Jaakkola, T. (2023). Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. International Conference on Learning Representations.

20. Sandbrink, J. B. (2023). Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. arXiv preprint arXiv:2306.13952. https://doi.org/10.48550/arXiv.2306.13952