Physics-Informed Generative Modeling for De Novo Protein Design Under Thermodynamic and Structural Constraints

Mahmut Jungfer
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
jungfer946@buffalo.edu

Hemi McMurray
Department of Computer Science, University of North Texas, Denton, TX, USA.
hemim@unt.edu

Kenan Birnbaum
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
kenanb@ku.edu

Abstract

De novo protein design has entered a new phase in which generative machine learning systems propose sequences and structures that are not directly observed in nature. These systems promise advances in therapeutics, biocatalysis, materials, and environmental sensing, but their outputs must satisfy thermodynamic stability, structural compatibility, functional specificity, and manufacturability constraints. This paper examines physics-informed generative modeling as a socio-technical infrastructure rather than a narrow algorithmic pursuit. It argues that thermodynamic and structural constraints should be embedded across data generation, model architecture, training objectives, evaluation pipelines, and deployment governance. The analysis considers trade-offs among all-atom fidelity, computational tractability, equivariance, robustness, interpretability, and equity of access. It also evaluates infrastructure requirements, including high-performance computing, curated structural databases, reproducible evaluation suites, model documentation, and sustainability accounting. Cross-domain comparisons with molecular design, weather modeling, and safety-critical machine learning clarify how failure modes propagate when physical plausibility is treated as an afterthought. The paper concludes that successful de novo protein design requires hybrid architectures that couple generative proposal mechanisms with physics-based validation, uncertainty quantification, and governance mechanisms that make limitations visible to downstream users. Without such integration, generative models may produce aesthetically plausible but thermodynamically fragile or functionally irrelevant designs. With it, physics-informed generative systems can expand the space of viable proteins while maintaining scientific accountability, operational robustness, and public trust.

Keywords

physics-informed generative modeling; de novo protein design; thermodynamic constraints; structural biology; machine learning systems; governance; sustainability; robustness.

1. Introduction

Protein design is one of the most consequential frontiers in computational biology because proteins mediate catalysis, signaling, immunity, transport, and material assembly. For decades, de novo design depended on energy functions, fragment assembly, and expert-guided mutagenesis. The classical principle that sequence encodes structure and structure determines function remains foundational [1], but the design problem reverses this logic by seeking sequences and structures that realize desired functions. Recent advances in structure prediction and generative modeling have made this reversal increasingly tractable. Deep learning systems can predict folded structures with high accuracy [2], generate protein backbones conditioned on motifs [3], design sequences for given backbones [4], and explore programmable protein spaces [5]. These capabilities shift the bottleneck from proposing candidate molecules to validating, filtering, manufacturing, and governing them.

The central challenge is that generative plausibility is not equivalent to thermodynamic viability. A model may produce a conformation that appears physically reasonable under a learned distribution while lacking sufficient stability under physiological conditions. It may satisfy backbone geometry but neglect side-chain packing, electrostatic complementarity, solvation, or entropic penalties. It may bind a target in silico while failing in vitro because of aggregation, expression toxicity, or off-target interactions. Consequently, physics-informed generative modeling must be understood as a system-level discipline that connects machine learning architectures, simulation methods, experimental feedback, data infrastructure, and institutional oversight. The design of the model is inseparable from the design of the validation ecosystem around it.

This paper advances three arguments. First, thermodynamic and structural constraints should be treated as first-class system requirements rather than post hoc filters. Second, architecture choices such as equivariance, all-atom representation, and multimodal conditioning create predictable trade-offs among fidelity, compute cost, generalization, and interpretability. Third, deployment of de novo protein design systems requires governance mechanisms that address reproducibility, fairness of access, dual-use risk, environmental cost, and liability. These arguments place generative protein design within the broader literature on machine learning systems, scientific infrastructure, and responsible innovation.

2. Foundations of Physics-Informed Generative Protein Design

Generative protein design draws from several lineages. Structure prediction models learn mappings from sequence to three-dimensional conformation, often using evolutionary information and geometric attention [2]. Inverse folding models learn the reverse mapping from structure to sequence and are central to designing sequences compatible with a target backbone [4, 6]. Diffusion and flow-based generative models learn to denoise or transport samples through a learned distribution, enabling conditional generation of backbones, motifs, and molecular geometries [3, 14]. Protein language models learn evolutionary regularities from large sequence corpora and can propose variants with desirable properties [15]. Molecular generative models have similarly explored inverse design in chemistry and materials, where desired properties guide the generation process [9, 10].

Physics-informed approaches add inductive biases that reflect physical symmetries and thermodynamic principles. Equivariant neural networks ensure that predictions transform consistently under rotation and translation, which is essential for molecular systems [8, 12]. Equivariant diffusion models extend this idea to generative processes in three dimensions [13]. Geometric vector perceptrons and related architectures encode local chemical environments while preserving structural relationships [11]. Boltzmann generators seek to sample

equilibrium states of many-body systems, connecting generative modeling directly to statistical mechanics [7]. These developments suggest that generative protein design can be framed as sampling from a conditional distribution shaped by energy landscapes, structural constraints, and functional objectives.

The required reference for this analysis describes a rationalized all-atom protein design framework that unifies multimodal Bayesian flow [16]. Its significance lies less in any single architectural novelty than in the integration of sequence, structure, and chemical detail within a coherent generative process. This aligns with the broader claim that de novo design must move beyond backbone-level generation toward all-atom reasoning. Yet all-atom fidelity increases computational cost and data requirements. It also raises validation burdens because small errors in bond lengths, torsions, or protonation states can cascade into larger functional failures. The system-level question is therefore how to allocate resources between high-fidelity generation and high-throughput screening.

Data resources underpin these models. The Protein Data Bank provides experimentally determined structures that anchor supervised learning and validation [18]. UniProt supplies sequence and functional annotation at scale [19]. Structure prediction systems such as AlphaFold have expanded the structural coverage available for training and hypothesis generation [2], while newer multimodal systems aim to predict interactions among proteins, ligands, and nucleic acids [17]. However, predicted structures are not experimental ground truth, and datasets carry historical biases toward well-studied protein families. Physics-informed modeling must therefore combine data-driven priors with simulation-based constraints and experimental feedback loops.

3. Thermodynamic and Structural Constraints in Generative Systems

Thermodynamic constraints enter protein design at multiple levels. Global stability depends on the balance between enthalpic interactions and entropic costs, including backbone conformational entropy and side-chain packing. Local structural constraints include bond geometry, chirality, steric exclusion, and hydrogen bonding. Functional constraints include active-site geometry, binding affinity, specificity, and allosteric regulation. A generative model that ignores these constraints may produce samples that are geometrically valid but thermodynamically implausible. A physics-informed model can incorporate them through energy-based scoring, differentiable simulators, constraint-aware sampling, or hybrid proposal mechanisms that alternate between learned generation and physical refinement.

Structural constraints are equally demanding because proteins are not static objects. They populate ensembles of conformations, and function often depends on flexibility, induced fit, or disorder. A design that appears rigid in a single predicted structure may be too flexible in solution or too constrained to bind a target. Generative systems must therefore represent uncertainty over conformations rather than collapsing to a single pose. This is difficult for models trained on static structures. It requires either ensemble data, molecular dynamics simulations, or probabilistic representations that capture conformational heterogeneity. The challenge is compounded when the target function involves transient interactions or multistep catalysis.

Thermodynamic and structural constraints also interact with manufacturability. A protein that is stable in simulation may be difficult to express in a host organism, fold correctly in the endoplasmic reticulum, avoid aggregation, or remain soluble at high concentration. These properties are not fully captured by folding free energy alone. They depend on cellular quality

control, codon usage, post-translational modification, and purification processes. A system-level approach to generative design must therefore integrate biophysical constraints with bioprocess constraints. Otherwise, the most elegant in silico candidates may fail at the earliest experimental stage.

The placement of physical constraints within the generative pipeline is a major design decision. Early integration shapes the model's representation and training objective, potentially improving sample efficiency and physical realism. Late integration preserves generative flexibility but may require extensive filtering and rejection sampling. Hybrid approaches use learned generators for exploration and physics-based evaluators for exploitation. Each choice affects computational cost, interpretability, and failure modes. For example, a model trained with strong equivariance may generalize better across orientations but may underrepresent rare structural motifs. A model with explicit all-atom energy terms may be more accurate but harder to scale. These trade-offs are not merely technical; they determine which proteins can be designed, by whom, and at what cost.

4. Architecture and System-Level Trade-Offs

Architecture choices in physics-informed generative protein design reflect competing priorities. Equivariant graph neural networks encode rotational and translational symmetry, making them well suited to molecular data [12]. Diffusion and flow models provide flexible generative processes that can be conditioned on structural motifs, sequences, or functional properties [13, 14]. Inverse folding architectures often focus on sequence recovery and structural compatibility, which makes them efficient but potentially limited when the desired function requires backbone flexibility [4, 6]. All-atom models attempt to capture chemical detail but require larger parameter spaces and more expensive sampling [16, 17]. The practical question is not which architecture is universally superior but which combination of representations and constraints matches the intended design task.

A recurring trade-off is between coarse-grained efficiency and all-atom fidelity. Coarse-grained models can explore large conformational spaces quickly and are useful for early-stage ideation. All-atom models can evaluate side-chain packing, hydrogen bonding, and electrostatic interactions more accurately but are computationally expensive. A tiered system may use coarse-grained generation to propose candidates, all-atom refinement to improve physical realism, and physics-based simulation to estimate stability. This tiered architecture resembles other engineering pipelines in which inexpensive models filter candidates before expensive evaluation. The risk is that errors introduced at the coarse stage may be impossible to repair later, especially if the backbone is topologically incompatible with the desired function.

Another trade-off concerns conditional control. Generative models can be conditioned on target binding sites, catalytic residues, symmetry constraints, or sequence motifs. Strong conditioning improves controllability but may reduce diversity. Weak conditioning preserves exploration but may produce many irrelevant samples. Physics-informed conditioning can guide generation toward low-energy states, but it may also bias the model toward known folds and away from novel topologies. This tension between novelty and plausibility is central to de novo design. A system that only generates known-like proteins may be safe but scientifically incremental; a system that generates radically novel folds may be exciting but difficult to validate.

System-level architecture must also address uncertainty and calibration. A generative model should communicate not only what it proposes but how confident it is. Uncertainty can arise from limited training data, ambiguous functional specifications, imperfect energy models, or stochastic sampling. If uncertainty is ignored, downstream users may treat generated designs as reliable candidates. If uncertainty is overestimated, useful designs may be discarded. Calibration methods, ensemble predictions, and physics-based consistency checks can help, but they add complexity. Governance frameworks such as model cards and datasheets provide structured ways to document intended use, limitations, and evaluation conditions [20, 21]. These instruments are especially important when generative outputs enter experimental or clinical pipelines.

5. Data, Infrastructure, and Deployment Pipelines

The infrastructure for de novo protein design includes curated databases, high-performance computing, simulation software, machine learning frameworks, laboratory automation, and experimental validation. The Protein Data Bank remains a central repository of structural information [18], while UniProt provides sequence and functional annotations [19]. Predicted structure databases expand coverage but introduce dependence on model versioning and confidence estimates [2]. Training pipelines must manage data provenance, leakage, and bias. If a model is evaluated on proteins similar to its training set, apparent performance may not transfer to novel folds or functions. Rigorous splits by fold, family, and function are therefore essential.

Deployment pipelines are often more complex than model development. A generated design must move through sequence optimization, expression testing, purification, stability assays, binding assays, and functional characterization. Each stage produces data that can feed back into the generative model, creating a closed-loop system. Closed-loop design promises faster iteration but requires robust data management, reproducible protocols, and statistical methods for dealing with small experimental batches. Without careful design, feedback loops can amplify biases or overfit to assay artifacts. The infrastructure must therefore support traceability from model version to generated candidate to experimental result.

Reproducibility is another infrastructure challenge. Generative models are sensitive to random seeds, data preprocessing, model checkpoints, and sampling hyperparameters. A design that cannot be reproduced cannot be reliably evaluated or improved. Containerization, version control, and standardized benchmarks can mitigate these problems, but they also require institutional investment. In academic settings, compute constraints may limit the scale of experiments, while in industrial settings, proprietary data and models may limit transparency. A balanced ecosystem might combine open benchmarks for fundamental methods with controlled access to sensitive or commercial datasets.

Deployment also raises questions of access and equity. If high-quality generative protein design requires expensive compute, proprietary databases, and automated laboratories, then the ability to design proteins may concentrate in a few institutions. This concentration could slow progress in neglected diseases, low-resource settings, and public-interest biotechnology. Open-source models, shared compute allocations, and international collaborations can broaden participation, but they must be paired with governance to prevent misuse. The design of infrastructure is therefore not neutral; it shapes who can ask which biological questions and who benefits from the answers.

6. Robustness, Fairness, and Governance

Robustness in generative protein design means maintaining performance under distribution shift, noisy data, adversarial inputs, and changing experimental conditions. A model trained on soluble globular proteins may fail on membrane proteins, intrinsically disordered regions, or engineered scaffolds. A model trained on one organism may not generalize to another. Robustness also includes resistance to reward hacking, where a generative model exploits weaknesses in a scoring function to produce high-scoring but nonfunctional designs. Physics-based validation and ensemble evaluation can reduce this risk, but no single metric is sufficient. Robustness must be assessed across multiple tasks, hosts, and environmental conditions.

Fairness in this context is not only about allocation of computational resources. It also concerns representation in datasets and priorities in application. Protein databases overrepresent proteins from well-studied organisms and disease areas, which can bias generative models toward certain folds and functions. If design efforts focus disproportionately on profitable therapeutics, neglected diseases and environmental applications may receive less attention. Fairness also applies to credit and benefit sharing when designs are derived from Indigenous knowledge, biodiversity, or community-collected samples. Governance mechanisms should therefore address data provenance, consent, and equitable access to resulting technologies.

Dual-use risk is a serious governance concern. Generative protein design could be used to create toxins, enhance pathogens, or bypass biosafety controls. The same capabilities that enable beneficial therapeutics can enable harmful applications. Governance must therefore combine technical safeguards, such as screening generated sequences against known hazards, with institutional oversight, export controls, and international norms. Overly restrictive governance could stifle legitimate research, while permissive governance could enable misuse. A layered approach is more promising: responsible disclosure, biosecurity review, audit trails, and clear red lines for dangerous experiments. Model cards and datasheets can support transparency, but they are not substitutes for enforceable policy [20, 21].

Governance must also address liability and accountability. If a generative model proposes a protein that causes harm in a clinical or environmental setting, who is responsible? The model developer, the user, the institution, or the regulator? Clear documentation of intended use, limitations, and validation status can help allocate responsibility. Regulatory frameworks for biotechnology, pharmaceuticals, and environmental release may need adaptation for AI-generated candidates. The challenge is to avoid both innovation paralysis and irresponsible deployment. Adaptive governance, which updates rules as evidence accumulates, may be more suitable than static regulation for a rapidly evolving field.

7. Sustainability and Policy Implications

Sustainability in generative protein design has at least two dimensions. The first is computational sustainability. Training and sampling from large generative models consume energy and hardware resources. The environmental cost of machine learning has been documented in natural language processing and broader AI systems [22]. Protein design may require molecular dynamics simulations, quantum chemistry calculations, and large-scale screening in addition to model training. These costs should be measured and reported. Efficient architectures, mixed-precision training, model distillation, and shared compute infrastructure can reduce environmental impact. However, efficiency gains may be offset by rebound effects if lower costs encourage ever-larger experiments.

The second dimension is biological and industrial sustainability. Protein design can contribute to sustainable chemistry, carbon capture, bioremediation, and biodegradable materials. It can also create new dependencies on proprietary platforms, patented sequences, and centralized manufacturing. Policy should encourage open scientific exchange while protecting legitimate intellectual property. Public funding agencies can require data sharing, reproducibility standards, and environmental impact statements for large computational projects. They can also support public-interest design challenges for neglected diseases and climate applications.

Policy implications extend to biosafety, biosecurity, and international coordination. National regulations may differ in their treatment of AI-generated biological sequences. If one country permits unrestricted generation while another restricts it, harmful actors may exploit regulatory gaps. International agreements, professional codes of conduct, and technical standards for sequence screening can help. At the same time, policy must avoid disproportionate burdens on low-resource researchers. Capacity building, shared reference laboratories, and open-source tools can promote equitable participation. Sustainability, security, and equity are therefore interconnected goals rather than separate concerns.

8. Future Directions

Future research should pursue tighter integration between generative models and physical simulation. Differentiable molecular dynamics, learned force fields, and hybrid quantum-classical methods could allow gradients from physics-based objectives to shape generative proposals. This would move beyond post hoc filtering toward true physics-informed generation. However, differentiable simulators remain expensive and unstable for large biomolecular systems. Progress will require advances in numerical methods, hardware acceleration, and uncertainty quantification. It will also require benchmarks that measure thermodynamic realism, not just structural similarity.

Another direction is multimodal conditioning that combines sequence, structure, function, and context. A protein's behavior depends on its environment, including pH, temperature, ionic strength, and binding partners. Generative models that condition on these variables could design proteins for specific industrial or physiological conditions. They could also design switchable proteins that respond to external stimuli. Such models would need training data that captures environmental variation, which is currently sparse. Experimental platforms that systematically vary conditions and measure responses could provide valuable data.

Finally, the field needs governance research that is as rigorous as the technical research. This includes methods for auditing generative models, measuring dual-use risk, evaluating fairness in access and application, and assessing environmental impact. Participatory approaches that involve biologists, engineers, ethicists, policymakers, and community stakeholders can help align design priorities with social values. The goal is not to slow innovation but to make its direction and consequences more deliberate. Physics-informed generative modeling will be most transformative when it is embedded in institutions capable of validating, regulating, and distributing its benefits responsibly.

9. Conclusion

Physics-informed generative modeling offers a powerful path toward de novo protein design under thermodynamic and structural constraints. By integrating physical symmetries, energy-based reasoning, structural databases, and multimodal generative processes, these systems can explore protein space more effectively than purely data-driven or purely simulation-based approaches. Yet the central challenges are system-level rather than merely architectural. They

include balancing all-atom fidelity with computational cost, embedding constraints throughout the pipeline, ensuring robustness under distribution shift, documenting limitations, governing dual-use risk, and distributing benefits equitably. The required all-atom multimodal Bayesian flow framework [16] illustrates the direction of travel, but no single model can resolve these broader tensions.

The most promising future lies in hybrid ecosystems that combine generative proposal, physics-based evaluation, experimental feedback, and transparent governance. Such ecosystems should treat thermodynamic viability, structural compatibility, manufacturability, and social responsibility as connected requirements. They should invest in shared data, reproducible benchmarks, sustainable compute, and inclusive participation. If these conditions are met, de novo protein design can move from impressive demonstrations to reliable, equitable, and scientifically accountable infrastructure. If they are neglected, the field risks producing brittle designs and concentrated benefits. The choice is not between physics and machine learning, nor between innovation and regulation, but among different ways of integrating them.

References

1. Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. Science, 181(4096), 223-230.

2. 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.

3. 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.

4. 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.

5. Ingraham, J. B., Baranov, M., Costello, Z., Barber, K. W., Wang, W., Ismail, A., ... Grigoryan, G. (2023). Illuminating protein space with a programmable generative model. Nature, 623(7989), 1070-1078.

6. Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., & Rives, A. (2022). Learning inverse folding from millions of predicted structures. Proceedings of the 39th International Conference on Machine Learning, 162, 8946-8970.

7. Noé, F., Olsson, S., Köhler, J., & Wu, H. (2019). Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. Science, 365(6457), eaaw1147.

8. Köhler, J., Klein, L., & Noé, F. (2020). Equivariant flows: Exact likelihood generative learning for symmetric densities. Proceedings of the 37th International Conference on Machine Learning, 119, 5361-5370.

9. Sanchez-Lengeling, B., & Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: Generative models for matter engineering. Science, 361(6400), 360-365.

10. Gebauer, N. W. A., Gastegger, M., Hessmann, S. S. P., Müller, K.-R., & Schütt, K. T. (2022). Inverse design of 3d molecular structures with conditional generative neural networks. Nature Communications, 13, 973.

11. Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L., & Dror, R. (2021). Learning from protein structure with geometric vector perceptrons. International Conference on Learning Representations.

12. Satorras, V. G., Hoogeboom, E., & Welling, M. (2021). E(n) equivariant graph neural networks. Proceedings of the 38th International Conference on Machine Learning, 139, 9323-9332.

13. Hoogeboom, E., Satorras, V. G., Vignac, C., & Welling, M. (2022). Equivariant diffusion for molecule generation in 3D. Proceedings of the 39th International Conference on Machine Learning, 162, 8867-8887.

14. 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.

15. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., ... Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637), 1123-1130.

16. 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.

17. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., ... Jumper, J. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 630(8016), 493-500.

18. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., ... Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Research, 28(1), 235-242.

19. UniProt Consortium. (2023). UniProt: The universal protein knowledgebase in 2023. Nucleic Acids Research, 51(D1), D523-D531.

20. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220-229.

21. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

22. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645-3650.