Uncertainty-Aware Generative Protein Design Through Multimodal Bayesian Flow Matching and Evolutionary Constraints

Mackenzie Stephenson
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
mackenziemail@ucf.edu

Arja Mikkonen
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
mikkonen1992@unr.edu

Dietlinde Killer
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
killer944@uc.edu

Abstract

Generative protein design increasingly combines deep learning with biophysical and evolutionary reasoning, yet deployed systems still struggle to represent uncertainty, satisfy manufacturability constraints, and remain accountable across scientific and industrial settings. This paper examines uncertainty-aware generative protein design through the lens of multimodal Bayesian flow matching and evolutionary constraints. Rather than treating protein generation as a single-model sequence prediction problem, the discussion frames it as a socio-technical infrastructure in which structure prediction, sequence design, all-atom refinement, evolutionary scoring, wet-lab validation, and governance processes interact. The paper argues that multimodal Bayesian flow matching offers a useful architectural middle ground between probabilistic modeling and scalable generative synthesis, while evolutionary constraints provide biologically grounded regularization that can reduce pathological outputs and improve deployability. At the same time, these advantages introduce trade-offs involving computational cost, calibration, data provenance, model plurality, and institutional responsibility. The analysis covers system architecture, deployment pipelines, robustness, fairness, sustainability, evaluation metrics, and policy implications. It emphasizes that uncertainty should not be treated only as a numerical confidence estimate but as an operational property distributed across data, models, laboratory workflows, and decision rights. The paper concludes that credible protein design platforms require hybrid generative architectures, evolutionary constraints, calibrated uncertainty reporting, and governance mechanisms that make model limitations visible to scientific and regulatory stakeholders. Such systems can accelerate discovery only if they preserve scientific traceability and avoid overstating the reliability of generated candidates.

Keywords

protein design, Bayesian flow matching, uncertainty quantification, evolutionary constraints, multimodal learning, AI governance, synthetic biology, robust deployment.

1. Introduction

Generative protein design has become a central frontier in computational biology and artificial intelligence because it promises to compress years of directed evolution and structural screening into computational search. Early breakthroughs in protein structure prediction demonstrated that deep neural networks could learn representations of molecular geometry and evolutionary covariation at unprecedented scale [1], [2]. Subsequent work on diffusion models, language models, and graph-based design systems expanded the field from prediction toward controllable generation [3], [4], [5]. These systems can propose sequences, backbones, motifs, and all-atom assemblies that satisfy functional objectives, but their outputs are not equally reliable across target families, experimental conditions, or downstream manufacturing contexts. The central challenge is therefore not merely generating plausible proteins but generating candidates whose uncertainty is characterized, whose evolutionary plausibility is respected, and whose failure modes are legible to scientists and decision-makers.

This paper explores that challenge by connecting three domains that are often treated separately. The first is multimodal generative modeling, in which sequence, structure, chemical features, and functional annotations are modeled jointly rather than through sequential handoffs. The second is Bayesian flow matching, a family of generative approaches that combines flexible transport-based generation with probabilistic inference and iterative refinement. The third is evolutionary constraint modeling, which uses phylogenetic, mutational, and fitness signals to keep generated candidates within biologically meaningful regions of design space. A recent unified multimodal Bayesian flow framework for all-atom protein design illustrates the potential of integrating these strands [17]. However, the purpose here is not to promote one architecture but to examine the systemic conditions under which uncertainty-aware generative design can be deployed responsibly.

The argument advanced here is that uncertainty-aware protein design should be understood as an infrastructure problem. A model that generates a candidate protein is only one component in a longer pipeline that includes target specification, data curation, model training, uncertainty estimation, evolutionary filtering, all-atom validation, laboratory synthesis, and regulatory review. Each stage contributes to the overall reliability of the system. If uncertainty is computed only at the sequence level but ignored during synthesis prioritization, then downstream decisions may treat fragile candidates as robust. If evolutionary constraints are applied as a post hoc filter, then the generator may waste capacity exploring impossible or maladaptive regions. If governance focuses only on model performance while ignoring data provenance and validation infrastructure, then accountability gaps will persist even when benchmark scores improve.

The discussion also responds to a broader concern in artificial intelligence research: high-performing generative models often obscure their epistemic boundaries [15], [16], [18]. Protein design magnifies this concern because the cost of a false positive can be measured in failed experiments, wasted reagents, delayed therapeutic programs, or unsafe biological constructs. Uncertainty-aware design therefore requires more than adding a variance estimate to a neural network. It requires architectural choices that preserve multiple hypotheses, evaluation practices that test calibration under distribution shift, and institutional routines that allow uncertainty to influence resource allocation and safety review. Evolutionary constraints provide one source of external discipline because they encode, however imperfectly, the accumulated signal of natural selection and functional viability.

The paper proceeds as follows. Section 2 establishes conceptual foundations and the limitations of single-model design. Section 3 examines multimodal Bayesian flow matching as a generative infrastructure. Section 4 analyzes evolutionary constraints and hybrid search. Section 5 discusses system architecture and deployment. Section 6 addresses robustness and validation. Section 7 considers fairness, governance, and policy. Section 8 examines sustainability and scalability. Section 9 provides cross-domain comparisons and case illustrations. Section 10 outlines future research directions, and Section 11 concludes.

2. Conceptual Foundations and Problem Setting

Protein design is a constrained search problem over an enormous combinatorial space. Sequence determines structure, but structure also depends on environmental conditions, binding partners, post-translational modifications, and dynamic ensembles. Function is similarly context-dependent, meaning that a candidate optimized for one assay may fail in another. Traditional computational design methods relied on physics-based energy functions, rotamer libraries, and heuristic search, with mixed success across target classes. Deep learning changed the scale and speed of search by learning empirical distributions from protein databases and experimental datasets [6], [7], [8]. Yet learned distributions can reproduce biases in the training data, overrepresent well-studied families, and generate sequences that appear plausible under a model but are not viable in a cell.

Uncertainty enters this setting in several forms. Aleatoric uncertainty arises from experimental noise, stochastic folding, and incomplete characterization of environmental conditions. Epistemic uncertainty arises from limited data, model misspecification, and ignorance about regions of sequence space that have not been sampled. Structural uncertainty concerns backbone flexibility, side-chain packing, and all-atom detail. Functional uncertainty concerns whether a designed protein will bind, catalyze, resist degradation, or avoid immunogenicity. These forms of uncertainty are not interchangeable. A model may be confident about local geometry while being uncertain about biological function. An uncertainty-aware system should therefore distinguish what it knows, what it predicts, and what it cannot resolve with available evidence.

Earlier generative approaches provide important context. Variational autoencoders offered latent-variable representations for protein sequences but often produced blurred or conservative samples [13]. Generative adversarial networks improved sample sharpness in some domains but introduced training instability and mode collapse [14]. Diffusion and flow-based models have since become dominant because they can generate high-quality samples through iterative denoising or transport [11], [12]. Protein-specific diffusion models demonstrated that geometric priors can guide backbone generation and motif scaffolding [3], [10]. However, most systems still separate generation from uncertainty estimation and evolutionary evaluation. This separation creates integration risk: a candidate may be generated with high confidence under one objective, then rejected by an evolutionary filter, leaving unclear whether the generator, the filter, or the objective specification was at fault.

Multimodal modeling offers a partial response by allowing sequence, structure, and chemical context to inform one another during training and inference. Protein language models capture evolutionary and functional patterns from large sequence corpora [6], [7]. Structure prediction networks capture geometric constraints with high accuracy [1], [2]. Design models can condition on motifs, folds, or binding sites [3], [4]. Yet multimodality also increases complexity. Different modalities have different noise profiles, missingness patterns, and scales of reliability. A system that treats all modalities as equally trustworthy may allow a

noisy annotation to dominate a structural signal. Bayesian flow matching is attractive because it provides a principled way to represent uncertainty during iterative generation, but it does not eliminate the need for data curation, external constraints, or governance.

The design problem is further complicated by the fact that protein engineering is not a one-shot prediction task. It is a sequential decision process involving synthesis, assay, learning, and redesign. Each round of experimentation provides new data that should update the model or the search policy. If uncertainty is ignored, the system may repeatedly propose candidates in regions where the model is overconfident but wrong. If uncertainty is used only for ranking, the system may fail to explore informative failures that could improve future rounds. Evolutionary constraints can help balance exploitation and exploration by identifying regions that are mutationally accessible and functionally plausible, but they must be calibrated against the specific organism, expression system, and application.

3. Multimodal Bayesian Flow Matching as Generative Infrastructure

Bayesian flow matching combines ideas from continuous normalizing flows, diffusion processes, and Bayesian inference. In conceptual terms, it learns a transport path from a simple prior distribution to a complex data distribution while maintaining a probabilistic interpretation of intermediate states. This makes it well suited to protein design because generation can be staged: coarse structural features, residue identities, side-chain conformations, and chemical properties can be refined progressively. Unlike a purely discriminative model that maps input to output, a flow-based generative model can represent multiple plausible candidates and update beliefs as constraints are introduced. When extended to multiple modalities, it can integrate sequence, structure, and annotation signals without collapsing them prematurely into a single representation.

The multimodal aspect is particularly important for all-atom design. Sequence-only models may generate plausible amino acid strings that fail to pack or fold. Backbone-only models may produce geometries that cannot accommodate functional side chains. All-atom models must account for steric clashes, hydrogen bonding, electrostatics, and solvent exposure. A Bayesian flow framework can treat these properties as coupled modalities with different levels of uncertainty. For example, a backbone may be relatively well determined while side-chain rotamers remain uncertain. A binding interface may be structurally plausible but functionally ambiguous because the model has limited data on similar complexes. By maintaining distributions over latent and observed variables, the system can propose candidates while exposing where refinement or experimental validation is most needed.

A recent unified multimodal Bayesian flow approach for all-atom protein design, cited here as [17], demonstrates how these principles can be operationalized. The significance of such work lies less in a single benchmark result than in the architectural claim that multimodal generation and uncertainty can be handled within one iterative framework. This claim aligns with the broader movement toward foundation models that integrate heterogeneous data sources [22]. However, integration also creates new failure modes. If one modality is corrupted, the flow may propagate error across modalities. If the prior is too weak, the model may generate physically invalid samples. If the prior is too strong, it may suppress novelty. The design of multimodal Bayesian flow systems therefore requires careful attention to modality weighting, missing data handling, and out-of-distribution detection.

From a system perspective, multimodal Bayesian flow matching should be deployed as a component within a larger orchestration layer. The orchestrator should specify objectives,

manage compute budgets, route candidates to evolutionary filters, invoke all-atom refinement, and record provenance. It should also decide when uncertainty is low enough to proceed to synthesis and when additional computation or human review is required. This orchestration role is analogous to hidden technical debt in machine learning systems, where interfaces, data dependencies, and configuration choices can dominate long-term reliability [21]. If the generative model is treated as a black box that outputs candidates, then uncertainty signals may be lost at the interface. If the orchestrator treats uncertainty as a first-class input, then the system can allocate resources adaptively.

Bayesian flow matching also raises questions about calibration and evaluation. A model may produce a distribution that is mathematically coherent but poorly calibrated to experimental outcomes. Calibration requires datasets that link predicted uncertainty to observed success rates across diverse protein families and assay conditions [15], [18]. Deep ensembles and Bayesian neural networks provide complementary tools for predictive uncertainty [16], [18], but they can be computationally expensive. Flow-based approaches may offer more efficient sampling, yet their uncertainty estimates must still be validated against wet-lab results. Without such validation, uncertainty becomes a rhetorical feature rather than an operational control.

Another architectural consideration is modularity. A monolithic multimodal model may be elegant but difficult to audit, update, or constrain. A modular system can separate sequence generation, structure refinement, evolutionary scoring, and safety screening. Each module can have its own uncertainty representation and version history. Modularity also supports substitution: if a new structure predictor becomes available, it can replace an older module without retraining the entire system. The trade-off is integration complexity and potential inconsistency between modules. Multimodal Bayesian flow matching can serve as a shared probabilistic backbone that reduces inconsistency while still permitting specialized components at the edges. The key is to define interfaces that preserve uncertainty and provenance rather than discarding them.

4. Evolutionary Constraints and Hybrid Search

Evolutionary constraints provide a biological counterweight to purely generative objectives. Natural protein families encode information about stability, folding, interactions, and functional tolerance that is difficult to capture from limited experimental data alone. Protein language models learn from evolutionary data implicitly [6], [7], but explicit constraints can be applied through conservation scores, co-variation patterns, mutational tolerance predictors, and phylogenetic priors. These constraints can be used during training, during generation, or as post hoc filters. Each placement has different implications. Training-time constraints can shape the generative distribution, but they may reduce diversity and complicate updates. Inference-time constraints can guide sampling toward viable regions, but they require careful calibration to avoid overconstraining the search. Post hoc filtering is simple but can waste computation and obscure the relationship between generation and feasibility.

A hybrid search architecture can combine Bayesian flow matching with evolutionary scoring in an iterative loop. The generative model proposes candidates, the evolutionary module scores them, and the orchestrator decides whether to refine, resample, or synthesize. This loop resembles directed evolution, where mutation and selection are repeated over generations. The difference is that computational selection can evaluate many more variants than wet-lab screening. However, computational selection is only as good as its proxies. A candidate that scores well on a conservation metric may still fail to express. A candidate that violates a

conservation metric may still be functional in a new context. Evolutionary constraints should therefore be treated as probabilistic evidence rather than hard rules, and their strength should be adjustable based on target family, experimental system, and risk tolerance.

The interaction between generative novelty and evolutionary plausibility is a central trade-off. If the model is too constrained, it will rediscover known proteins and offer little advantage over natural scaffolds. If it is too unconstrained, it may generate sequences that are biologically nonsensical or unsafe. Multimodal Bayesian flow matching can help by representing uncertainty over evolutionary compatibility. For example, the model may generate a candidate with high confidence in local structure but low confidence in mutational tolerance. The orchestrator can then prioritize experimental assays that resolve that uncertainty. In this way, evolutionary constraints become part of an active learning strategy rather than a static filter.

Case illustrations from protein engineering show that successful designs often require a balance between rational mutation and exploratory search. Antibody engineering, enzyme redesign, and binder design each have different tolerance landscapes [5], [8]. Enzymes may tolerate mutations far from the active site but be highly sensitive near catalytic residues. Antibodies may require specific complementarity-determining region profiles while allowing variation in framework regions. Binders may need to satisfy interface shape complementarity while avoiding aggregation-prone patches. A single global evolutionary constraint cannot capture these local differences. A more sophisticated system should apply constraints at the level of domains, motifs, and residue neighborhoods, with uncertainty estimates for each region.

Evolutionary constraints also introduce governance questions. If a system uses pathogenicity predictions, conservation scores, or biosecurity filters, those components must be auditable and regularly updated. Constraints based on historical data may encode past sampling biases, such as overrepresentation of model organisms or well-funded research areas. They may also be used to justify decisions that are difficult to contest. A responsible system should document which constraints were applied, why they were chosen, and how they affect candidate ranking. It should allow human reviewers to override constraints when scientific justification warrants, while preserving a record of the decision. This is especially important in dual-use research, where safety and openness must be balanced [19], [20].

5. System Architecture and Deployment Considerations

A deployable uncertainty-aware protein design platform requires a layered architecture. The data layer manages sequence databases, structure repositories, experimental assays, and metadata about provenance and licensing. The model layer hosts multimodal Bayesian flow models, structure predictors, evolutionary scorers, and safety classifiers. The orchestration layer manages workflows, compute allocation, uncertainty thresholds, and human-in-the-loop review. The validation layer connects computational predictions to laboratory information management systems, synthesis vendors, and assay results. The governance layer defines access controls, audit trails, and compliance procedures. Each layer has distinct failure modes, and overall system reliability depends on how well they communicate.

Deployment environments vary widely. Academic laboratories may prioritize flexibility and low cost, industry teams may require reproducibility and intellectual property controls, and public health organizations may demand biosecurity oversight. A cloud-native deployment can scale compute but raises data sovereignty and privacy concerns. An on-premises

deployment can protect sensitive data but may limit access to state-of-the-art models. A hybrid model can balance these concerns, but it increases integration complexity. Multimodal Bayesian flow matching is computationally intensive, especially when all-atom refinement and uncertainty estimation are performed together. Deployment must therefore consider batching, caching, surrogate models, and early stopping. These optimizations should not compromise uncertainty reporting; if approximations are used, their effect on calibration should be measured.

Human-computer interaction is another critical dimension. Scientists need interfaces that display uncertainty in actionable ways. A single confidence score is insufficient. Reviewers need to see which regions of a candidate are uncertain, which constraints were active, which data were influential, and which experiments would most reduce uncertainty. Visual analytics can help, but they must avoid creating false precision. A heat map of per-residue uncertainty may be useful, but it can also imply that uncertainty is localized when it is actually systemic. The interface should encourage appropriate trust rather than blind acceptance. It should also support documentation so that computational predictions can be compared with experimental outcomes over time.

Operational metrics should go beyond top-k accuracy or structural similarity. A production system should track synthesis success rate, expression yield, binding affinity, stability, and safety screen outcomes. It should also track uncertainty calibration, coverage of design space, diversity of candidates, and time to result. These metrics are often in tension. A system that optimizes for synthesis success may become conservative and miss novel designs. A system that optimizes for novelty may generate candidates that are difficult to manufacture. Governance should define acceptable trade-offs based on application. For therapeutic proteins, safety and manufacturability may dominate. For research reagents, novelty and dynamic range may be more important.

Deployment also requires version control and reproducibility. Protein design models are updated frequently, and a candidate generated by one model version may not be reproducible by another. Provenance should include model weights, training data snapshots, random seeds, constraint settings, and software dependencies. This level of documentation is demanding, but it is necessary for scientific reproducibility and regulatory review. It also supports organizational learning: when a design fails, teams can trace whether the failure originated in data, model, constraints, or experimental conditions. Without such traceability, uncertainty-aware design becomes difficult to improve systematically.

6. Uncertainty, Robustness, and Validation

Uncertainty quantification is valuable only if it changes decisions. In protein design, this means using uncertainty to prioritize experiments, allocate compute, and trigger human review. A candidate with high predicted binding affinity but high uncertainty may be a better experimental subject than a candidate with slightly lower affinity but high confidence, because the former provides more information. This information-theoretic perspective aligns with active learning and Bayesian experimental design. However, it requires the system to estimate not only predictive uncertainty but also the expected value of information. That estimate depends on assay cost, turnaround time, and strategic priorities. A purely automated system may struggle to incorporate these factors, which is why human oversight remains essential.

Robustness testing should include distribution shift. Models trained on natural proteins may perform poorly on designed sequences far from natural distributions. Models trained on one organism may fail in another. Models trained on soluble proteins may fail on membrane proteins. Robustness also includes adversarial or accidental failure modes, such as sequences that satisfy computational objectives but violate safety constraints. Red-teaming and stress testing can help identify these cases. In biosecurity, adversarial misuse is a serious concern [19]. A responsible platform should include screening, access controls, and monitoring. These measures should be proportionate and transparent, not hidden behind proprietary claims.

Validation must be multi-level. In silico validation can check geometry, packing, and energetic plausibility. Evolutionary validation can check conservation and mutational tolerance. Experimental validation can measure expression, stability, binding, and function. Each level has different error rates and biases. A candidate that passes in silico validation may fail experimentally. A candidate that fails one computational filter may still succeed if the filter is miscalibrated. The system should therefore avoid hard cutoffs unless they are justified by safety or regulatory requirements. Instead, it should use probabilistic thresholds and document the rationale for each decision.

Calibration is especially challenging because experimental data are scarce and heterogeneous. Different assays have different noise levels, and negative results are often underreported. A model may appear well calibrated on a benchmark but poorly calibrated on a new target. Ongoing calibration requires prospective validation and feedback loops. This is an infrastructure challenge, not just an algorithmic one. It requires data sharing agreements, standardized metadata, and incentives for reporting failures. Without such infrastructure, uncertainty estimates will remain unreliable, and users may either ignore them or overtrust them.

7. Fairness, Governance, and Policy

Fairness in protein design may seem less obvious than in criminal justice or hiring, but it is still relevant. Design models can prioritize well-studied protein families, commercially valuable targets, or populations represented in biomedical datasets. They can also concentrate benefits in institutions with advanced compute and laboratory infrastructure. Governance should consider equitable access to design tools, data, and validation capacity. It should also consider how intellectual property and licensing affect downstream use. If generative platforms are controlled by a few organizations, the scientific community may lose the ability to audit or extend them. Open science initiatives can help, but they must be balanced with biosecurity and privacy.

Policy frameworks for artificial intelligence emphasize transparency, accountability, and human oversight [20]. Protein design adds biological safety, environmental release, and dual-use concerns. Regulatory agencies may require evidence of safety and efficacy, but they may lack expertise in generative model uncertainty. This creates a need for interdisciplinary governance bodies that include computational scientists, biologists, ethicists, security experts, and regulators. Such bodies should develop standards for uncertainty reporting, data provenance, and model documentation. They should also define when a designed protein requires additional containment or review. These standards should be adaptive because the technology is evolving rapidly.

Governance should not be reduced to compliance checklists. It should shape system design. For example, if auditability is a governance goal, then the architecture should log uncertainty,

constraints, and human decisions. If equitable access is a goal, then deployment should include tiered compute access or shared validation facilities. If safety is a goal, then screening should be integrated into the generation loop rather than added after synthesis. These choices are technical and political. Treating them as purely technical obscures value judgments about risk, benefit, and responsibility.

Policy also affects international collaboration. Protein design is global, and harmful constructs do not respect borders. Harmonized standards for biosecurity screening, data sharing, and export controls could reduce risk, but they must not stifle legitimate research. The challenge is to create governance that is robust without being brittle. This requires iterative learning, transparency about incidents, and mechanisms for updating rules as evidence accumulates. It also requires avoiding the false choice between open science and security. Both are important, and the goal is to design institutions that support both.

8. Sustainability and Scalability

Generative protein design is computationally expensive. Training large multimodal models requires significant energy, water, and hardware. Inference for all-atom refinement and uncertainty estimation can also be intensive. Sustainability concerns are often overlooked in AI research, but they become material at scale [21]. A platform that generates millions of candidates may consume substantial resources before a single experiment is run. Life-cycle assessment should include model training, data storage, inference, synthesis, and experimental validation. In some cases, computational screening may reduce overall resource use by avoiding failed experiments. In other cases, it may simply shift resource consumption upstream. The net effect depends on how candidates are prioritized and how many are synthesized.

Scalability requires algorithmic efficiency and infrastructure planning. Bayesian flow matching can be optimized through distillation, quantization, and adaptive computation. Evolutionary constraints can be approximated with fast surrogate models and then verified with more accurate methods. Multimodal fusion can be staged so that expensive modalities are invoked only when needed. These optimizations should be evaluated for their impact on uncertainty calibration. A faster model that loses uncertainty fidelity may be less sustainable overall if it leads to more failed experiments.

Scalability also has social dimensions. If only well-resourced institutions can run advanced design platforms, the benefits of protein engineering will be unevenly distributed. Federated learning and shared model hubs can broaden access, but they raise governance challenges. Data contributors may not want their proprietary sequences used for model training. Model developers may not want their weights distributed without control. Sustainability therefore includes institutional sustainability: funding models, maintenance, and community support. A platform that depends on a single grant or company may disappear, leaving users without updates or reproducibility. Long-term stewardship requires diversified funding and transparent governance.

9. Cross-Domain Comparisons and Case Illustrations

Comparing protein design with other generative domains clarifies both opportunities and limits. In image generation, uncertainty may affect aesthetic choice or safety filtering. In protein design, uncertainty affects physical experiments with material and biological consequences. In drug discovery, generative chemistry shares concerns about synthesizability, toxicity, and patentability. Protein design adds the complexity of folding, expression, and

evolutionary compatibility. In materials science, generative models must respect thermodynamic and mechanical constraints. Protein design similarly requires physical and biological constraints, but its search space is shaped by billions of years of evolution. This makes evolutionary constraints unusually informative and unusually biased.

A case illustration can be found in binder design. A generative model proposes a scaffold and interface sequence, a structure predictor evaluates complementarity, an evolutionary module checks conservation of interface residues, and an all-atom refinement step resolves clashes. Uncertainty may be high at the interface because small changes can disrupt binding. The orchestrator might prioritize candidates with moderate predicted affinity but low structural uncertainty for experimental testing. If the first round fails, the results update the model and guide the next round. This iterative process resembles Bayesian optimization, where exploration and exploitation are balanced. The difference is that the objective function is noisy, expensive, and partially unknown.

Another case is enzyme redesign. Here, catalytic residues are highly constrained, while distal mutations can affect stability and substrate specificity. A multimodal model can treat the active site as a high-confidence region and the periphery as more uncertain. Evolutionary constraints can preserve catalytic motifs while allowing exploration elsewhere. All-atom refinement can check substrate binding and transition-state complementarity. Governance may require additional review if the enzyme could produce toxic compounds. The system should therefore integrate safety screening early, not as an afterthought.

These cases show that uncertainty-aware generative design is not a single algorithm but a workflow. The value of multimodal Bayesian flow matching lies in its ability to represent multiple hypotheses and refine them under constraints. The value of evolutionary constraints lies in grounding the search in biological plausibility. The value of governance lies in ensuring that decisions are transparent and accountable. None of these components is sufficient alone.

10. Future Research Directions

Future research should focus on several interconnected areas. First, uncertainty representation must become more compositional. A single global uncertainty score is insufficient for proteins with modular domains, flexible loops, and context-dependent interfaces. Models should estimate uncertainty at the level of residues, motifs, domains, and interactions. Second, calibration must be evaluated prospectively in wet-lab settings. This requires standardized benchmarks that include negative results and diverse protein families. Third, evolutionary constraints should be integrated adaptively rather than statically. The system should learn which constraints are most informative for a given target and adjust their strength over time.

Fourth, system architecture should support human-AI collaboration. Interfaces should make uncertainty, provenance, and constraint activity visible without overwhelming users. They should also support contestability, allowing scientists to challenge model outputs and record their reasoning. Fifth, governance should develop standards for auditability and safety screening. These standards should be technically informed and internationally coordinated. Sixth, sustainability should be treated as a design objective alongside accuracy and speed. This includes measuring the environmental cost of training, inference, and experimental validation.

Finally, research should examine the long-term epistemic effects of generative design. If scientists increasingly rely on models that propose candidates, will they lose the ability to

reason about protein structure and evolution? Or will models augment human expertise by handling routine search while leaving creative and critical thinking to researchers? The answer depends on how systems are designed and governed. A well-designed platform should support understanding, not replace it. It should expose assumptions, invite scrutiny, and learn from failures. These are not only technical goals but institutional and cultural ones.

11. Conclusion

Uncertainty-aware generative protein design through multimodal Bayesian flow matching and evolutionary constraints represents a promising direction for computational biology. It combines flexible generative modeling with probabilistic inference and biological grounding, offering a way to propose candidates while acknowledging what is unknown. Yet the success of such systems depends less on any single model than on the infrastructure that surrounds them. Data provenance, calibration, evolutionary scoring, all-atom refinement, human oversight, safety screening, and governance must be integrated into the design loop. Without this integration, uncertainty estimates may be ignored, evolutionary constraints may be misapplied, and generated proteins may fail in ways that are difficult to diagnose.

The paper has argued for a system-level perspective that treats protein design as a socio-technical infrastructure. Multimodal Bayesian flow matching can serve as a shared probabilistic backbone, but it must be modular enough to support auditing and substitution. Evolutionary constraints can provide biological discipline, but they must be probabilistic and context-sensitive. Governance can ensure accountability, but it must be adaptive and technically informed. Sustainability requires attention to compute, experiments, and institutional stewardship. These dimensions are not optional additions to an otherwise technical problem. They are constitutive of whether generative protein design can be trusted and useful at scale.

A recent unified multimodal Bayesian flow framework for all-atom design, cited as [17], illustrates the architectural possibilities, but the broader lesson is that uncertainty must be operationalized across the entire pipeline. The field should move toward systems that generate candidates, quantify uncertainty, apply evolutionary constraints, prioritize experiments, and learn from outcomes in a transparent loop. Such systems can accelerate discovery while preserving scientific rigor and public trust. The challenge is both technical and institutional, and it requires collaboration across artificial intelligence, biology, engineering, ethics, and policy. If met, uncertainty-aware generative protein design could become a durable foundation for responsible innovation.

References

1. 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., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.

2. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., ... Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science, 373(6557), 871-876.

3. 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., ... 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., 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., ... Baker, D. (2022). Robust deep learning-based protein sequence design using ProteinMPNN. Science, 378(6615), 49-56.

5. Hie, B. L., Shapero, K. K., Xu, Y., Shen, Z., Shanker, V. R., Yang, K. K., Lewis, S., Biddle, C., Chen, A., Chen, A., ... Kim, P. S. (2024). Efficient evolution of human antibodies from general protein language models. Nature Biotechnology, 42, 275-283.

6. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, 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.

7. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637), 1123-1130.

8. Madani, A., Krause, B., Greene, E. R., Subramanian, S., Mohr, B. P., Holton, J. M., Olmos, J. L., Xiong, C., Sun, Z. Z., Socher, R., Fraser, J. S., & Naik, N. (2023). Large language models generate functional protein sequences across diverse families. Nature Biotechnology, 41(8), 1099-1106.

9. Anishchenko, I., Pellock, S. J., Chidyausiku, T. M., Ramelot, T. A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A. K., DiMaio, F., Carter, L., Chow, C. M., Montelione, G. T., & Baker, D. (2021). De novo protein design by deep network hallucination. Nature, 600(7889), 547-552.

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

11. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840-6851.

12. Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., & Le, M. (2023). Flow matching for generative modeling. International Conference on Learning Representations.

13. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. International Conference on Learning Representations.

14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27, 2672-2680.

15. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, 70, 1321-1330.

16. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30, 5574-5584.

17. 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.

18. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30, 6402-6413.

19. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

20. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399.

21. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 28, 2503-2511.

22. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.