Multi-Objective Generative Optimization of Protein–Ligand Binding Affinity Using Bayesian Inference and Molecular Simulation

Ignacy Jantos
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
ignacywork@uab.edu

Abstract

The optimization of protein–ligand binding affinity is a central problem in drug discovery, chemical biology, and protein engineering, yet it remains difficult because affinity is governed by coupled thermodynamic, kinetic, and conformational factors that cannot be reduced to a single static score. Recent advances in generative artificial intelligence, Bayesian uncertainty modeling, and molecular simulation have created an opportunity to treat affinity optimization not as a narrow prediction task but as a multi-objective socio-technical system spanning data generation, model architecture, simulation fidelity, laboratory validation, and governance. This paper develops a system-level account of multi-objective generative optimization for protein–ligand binding affinity using Bayesian inference and molecular simulation. It argues that the most consequential challenges are not solely algorithmic but architectural and institutional: integrating heterogeneous data sources, calibrating uncertainty across simulation and assay scales, managing trade-offs among potency, selectivity, safety, synthesizability, and sustainability, and deploying optimization loops in ways that are reproducible, fair, and robust. The discussion connects molecular generative models, equivariant graph representations, diffusion and flow-based generation, free-energy simulation, active learning, and human-in-the-loop decision-making. It also examines infrastructure requirements, including high-performance computing, federated data governance, automated laboratories, audit trails, and model documentation. Through cross-domain comparisons with materials discovery and protein design, the paper identifies design principles for trustworthy generative optimization systems. It concludes that progress will depend on coupling Bayesian decision theory with physical simulation and on building governance structures that make uncertainty, trade-offs, and limitations explicit to researchers, regulators, and affected communities.

Keywords

protein–ligand binding affinity, generative optimization, Bayesian inference, molecular simulation, multi-objective optimization, uncertainty quantification, drug discovery, socio-technical systems.

1. Introduction

Protein–ligand binding affinity is a foundational quantity in molecular science because it shapes drug efficacy, selectivity, toxicity, and the feasibility of chemical probes. The prediction and optimization of affinity have traditionally relied on structure-based docking, molecular dynamics, free-energy perturbation, quantitative structure–activity relationships, and medicinal chemistry intuition. Each approach offers valuable information, but each also

carries limitations in accuracy, cost, transferability, or interpretability. Deep learning has altered this landscape by enabling rapid surrogate prediction and generative design, with protein structure prediction systems demonstrating how large-scale data and neural architectures can reshape molecular modeling [1], and earlier deep learning potentials showing the value of learned representations for protein structure [2]. However, the translation from accurate prediction to reliable optimization remains difficult because affinity is not a fixed property of a single molecular pose. It emerges from ensembles, solvation, entropy, electrostatic complementarity, conformational strain, and kinetic accessibility.

Generative optimization introduces a further layer of complexity. Instead of scoring a fixed library, generative models propose novel molecules or protein variants that are intended to satisfy multiple objectives simultaneously. This shifts the problem from ranking to search, and from prediction to decision-making under uncertainty. A system that optimizes only predicted affinity may produce molecules that are insoluble, toxic, difficult to synthesize, or inactive in a biological context. Consequently, contemporary work increasingly frames molecular design as a multi-objective problem involving potency, selectivity, pharmacokinetics, safety, synthetic accessibility, and environmental impact. The movement toward closed-loop discovery has been illustrated by deep learning approaches to antibiotic discovery [3] and rapid identification of kinase inhibitors [4], but such successes also reveal the infrastructure burden of validation, reproducibility, and governance.

This paper argues that multi-objective generative optimization of protein–ligand binding affinity should be understood as a socio-technical system rather than a standalone model. Bayesian inference provides a principled way to represent uncertainty, update beliefs after simulation or assay feedback, and guide acquisition of new data. Molecular simulation provides physical grounding that can constrain generative proposals and expose failure modes that purely data-driven models may miss. Generative models provide the search capacity to explore vast chemical and conformational spaces. Yet the integration of these components requires architectural choices about fidelity, latency, modularity, and human oversight. It also requires policy choices about data sharing, intellectual property, dual-use risk, and equitable access to benefits. The central claim is that durable progress will come from designing optimization loops that are uncertainty-aware, physically grounded, multi-objective, auditable, and socially accountable.

2. System-Level Foundations of Protein–Ligand Affinity Optimization

Binding affinity is often described through equilibrium dissociation constants, but its practical meaning depends on the biological question. In drug discovery, affinity must be considered alongside residence time, target engagement, off-target effects, and cellular context. In protein engineering, affinity may be coupled to stability, expression, and allosteric regulation. In chemical biology, affinity must be interpreted relative to selectivity across homologous proteins. These distinctions matter because a generative system optimized for one definition may perform poorly under another. A molecule that binds tightly in an isolated assay may fail in a cellular environment due to permeability, efflux, metabolism, or compartmentalization. Therefore, affinity optimization is not merely a regression problem; it is a problem of aligning molecular properties with a specified use context.

From a system perspective, the optimization pipeline includes data acquisition, representation, model training, candidate generation, simulation-based evaluation, experimental validation, and feedback. Each stage introduces uncertainty and bias. Public databases contain heterogeneous assay conditions, inconsistent units, and missing metadata. Proprietary datasets

may be larger but are rarely accessible for independent validation. Simulation outputs depend on force fields, sampling times, protonation states, and solvent models. Wet-lab measurements vary across instruments, protocols, and laboratories. A robust system must therefore track provenance and uncertainty across all stages, not only within the predictive model. Otherwise, Bayesian confidence may be miscalibrated precisely where decisions matter most.

The infrastructure surrounding this pipeline is equally important. Generative optimization requires compute for training and sampling, storage for molecular trajectories and assay records, orchestration for simulation workflows, and laboratory automation for synthesis and testing. It also requires human expertise in medicinal chemistry, structural biology, toxicology, and regulatory science. Treating the system as purely computational underestimates the coordination costs and the risk of automating flawed assumptions. A system-level view encourages modularity, so that improved simulation methods or new assay data can be integrated without rebuilding the entire pipeline. It also encourages explicit interfaces between models, simulations, and experimentalists, allowing each component to communicate uncertainty and constraints rather than only point predictions.

3. Bayesian Inference for Uncertainty-Aware Molecular Optimization

Bayesian inference offers a coherent framework for updating beliefs as new molecular, simulation, and experimental evidence becomes available. Rather than treating binding affinity as a deterministic label, a Bayesian model represents a distribution over plausible affinity values and parameters. This is important because molecular datasets are often small, noisy, and biased toward chemical series that were historically prioritized. A Bayesian workflow can distinguish aleatoric uncertainty, which arises from measurement noise, from epistemic uncertainty, which arises from limited knowledge. In drug discovery, this distinction affects decisions: high aleatoric uncertainty may call for more precise assays, whereas high epistemic uncertainty may call for exploring new chemical regions or protein conformations.

Generative modeling and Bayesian inference have complementary roles. Variational autoencoders and related latent-variable models provide probabilistic representations of molecular structure [6], while stochastic backpropagation and approximate inference enable training with latent uncertainty [13]. Graph neural networks and molecular graph convolutions extend these ideas to discrete molecular topologies [8], [9], and equivariant architectures improve physical consistency under rotation and translation [10]. Diffusion and score-based models offer powerful generative priors over molecular geometries and conformations [11], [12]. Recent work on unified multi-modal Bayesian flow for rationalized all-atom protein design [7] illustrates how Bayesian generative processes can be coupled with all-atom representations, suggesting a broader direction for affinity optimization that integrates sequence, structure, and ligand context rather than treating them as separate tasks.

Uncertainty-aware optimization also requires acquisition functions that balance exploration and exploitation. In active learning, the system selects simulations or experiments that are expected to improve the model or identify promising candidates. This is not a purely technical choice because acquisition criteria encode institutional priorities. If the system favors predicted potency, it may neglect diverse scaffolds or underrepresented targets. If it favors uncertainty reduction, it may spend resources on questions that are scientifically interesting but not therapeutically valuable. A well-designed Bayesian optimization loop therefore makes its preferences explicit and allows stakeholders to adjust them. Calibration, sensitivity

analysis, and posterior predictive checks become governance tools as much as statistical diagnostics.

4. Molecular Simulation as Grounding and Constraint Layer

Molecular simulation provides the physical grounding that generative models often lack. Docking can rapidly estimate binding poses, but it approximates flexibility and solvation. Molecular dynamics can capture conformational ensembles and solvent effects, but requires sufficient sampling and careful force-field selection. Alchemical free-energy methods can provide more rigorous affinity estimates, but their cost limits throughput. Enhanced sampling, Markov state models, and machine-learned potentials can extend accessible time scales, yet they introduce new approximations. The value of simulation in a generative optimization system is not only to score candidates but also to constrain the search space, reject physically implausible proposals, and reveal mechanisms that can guide subsequent design.

A multi-fidelity simulation architecture is often more practical than a single high-accuracy method. Low-fidelity docking or fast surrogates can filter large generative outputs, medium-fidelity molecular dynamics can refine promising poses, and high-fidelity free-energy calculations can be reserved for finalists. Bayesian inference can combine information across fidelities by modeling systematic discrepancies between methods. This approach recognizes that simulation and experiment are not interchangeable; they have different error structures and different costs. A system that treats simulation as ground truth may overfit to force-field artifacts, while a system that treats simulation as mere features may miss critical physical constraints. The productive posture is to treat simulation as a calibrated, uncertainty-bearing source of evidence.

Simulation also plays a role in robustness. A generative model trained on static protein structures may fail when the target undergoes conformational change, allosteric modulation, or mutation. Simulation can probe alternative states and expose sensitivity to protonation, water networks, and loop dynamics. In protein–ligand optimization, this is especially important because cryptic pockets and induced-fit effects can determine whether a molecule binds. By integrating simulation into the generative loop, the system can search not only over ligands but also over target conformations and environmental conditions. This expands the optimization space but also increases computational cost, making Bayesian scheduling and resource allocation essential.

5. Multi-Objective Generative Optimization

Multi-objective optimization recognizes that binding affinity is only one objective among many. Potency must be balanced with selectivity, solubility, permeability, metabolic stability, toxicity, synthesizability, and cost. In protein design, objectives may include thermostability, expression yield, and immunogenicity. In drug discovery, objectives may include target engagement in vivo and safety margins. These objectives often conflict. Increasing lipophilicity may improve membrane permeability but reduce solubility. Adding polar groups may improve solubility but reduce binding. Rigidifying a scaffold may improve affinity but complicate synthesis. A generative system must therefore navigate trade-offs rather than maximize a single scalar score.

Generative models provide several mechanisms for multi-objective search. Latent-variable models can encode molecular properties in continuous spaces, enabling gradient-based optimization and Bayesian sampling [5], [6], [13]. Graph-based and equivariant models can operate directly on molecular topology and geometry [8], [9], [10]. Diffusion and flow-based

models can generate diverse candidates while respecting geometric constraints [11], [12]. Inverse design approaches in materials and molecules demonstrate how generative models can propose candidates with desired properties [20]. The challenge is not the absence of generative architectures but the integration of objectives, constraints, and uncertainty into a coherent decision process.

A system-level treatment of multi-objective optimization must also address preference elicitation and conflict resolution. Pareto-based methods can identify non-dominated candidates, but they do not by themselves choose among them. Scalarization methods can simplify decisions but may hide trade-offs. Interactive optimization allows medicinal chemists and biologists to adjust priorities as knowledge evolves. Bayesian optimization can incorporate preferences through acquisition functions or utility models. Governance enters when objectives include safety, environmental impact, or equitable access, because these values cannot be reduced to technical metrics without normative choices. A mature system makes trade-offs visible, documents why certain candidates were selected or rejected, and allows revision as new evidence arrives.

6. Architecture and Deployment Infrastructure

The architecture of a generative optimization platform determines whether it can be reproduced, audited, and improved. A layered design is useful. The data layer manages molecular structures, assay results, simulation trajectories, and metadata. The representation layer converts these data into graph, sequence, geometric, or multi-modal encodings. The generative layer proposes candidates. The simulation layer evaluates candidates at multiple fidelities. The decision layer combines predictions, uncertainty, and preferences. The validation layer coordinates experimental testing and returns results. The governance layer oversees access, safety, documentation, and compliance. These layers must communicate through well-defined interfaces so that components can be replaced without destabilizing the whole.

Deployment infrastructure includes high-performance computing, cloud resources, workflow engines, and laboratory information management systems. Generative models and molecular simulations are computationally intensive, and their resource demands may conflict with sustainability goals. Reproducibility requires versioned datasets, model checkpoints, random seeds, simulation parameters, and containerized environments. Audit trails are necessary for regulatory submissions and for internal quality control. Security and privacy matter when proprietary compounds, patient-derived data, or dual-use sequences are involved. Federated learning and secure computation can enable collaboration without centralizing sensitive data, but they introduce coordination overhead and may exacerbate inequities if only well-resourced institutions can participate.

Human-in-the-loop design is not a temporary concession to immature automation; it is a durable requirement. Chemists, biologists, pharmacologists, and clinicians bring contextual knowledge that may not be present in training data. They can recognize implausible mechanisms, synthetic challenges, or safety concerns that a model may overlook. Automated laboratories can accelerate feedback, but they also require robust scheduling, error handling, and quality control. The architecture should therefore support mixed-initiative workflows in which humans set objectives, inspect uncertainty, and approve high-stakes decisions, while computational agents handle large-scale search and routine evaluation. This aligns with broader arguments that high-stakes machine learning should favor interpretability and accountability rather than opaque automation [14], [15].

7. Robustness, Fairness, Sustainability, and Policy

Robustness in affinity optimization means maintaining performance under distribution shift. Chemical space is open-ended, protein targets mutate, assay conditions change, and biological contexts vary. A model trained on one target family may fail on another. A generative model may exploit dataset artifacts, such as assay-specific scaffolds, rather than learning transferable binding principles. Robustness testing should include scaffold splits, temporal splits, mutation perturbations, and cross-target validation. Simulation can help by testing candidates under alternative protonation states, conformational ensembles, or solvent conditions. Bayesian uncertainty can flag when predictions are unreliable, but only if the model is calibrated and the deployment context respects its limitations.

Fairness and equity are often neglected in molecular optimization, yet they are central to health and environmental policy. If generative drug discovery focuses only on profitable markets, neglected diseases may remain underserved. If data and compute are concentrated in a few institutions, benefits may accrue unevenly. If models are trained predominantly on European or North American assay data, generalization to other populations or environments may suffer. Fairness also applies to computational access: open benchmarks, shared models, and capacity-building can broaden participation. However, openness must be balanced against dual-use risks and intellectual property. Governance frameworks should therefore combine transparency, controlled access, and benefit-sharing mechanisms.

Sustainability is another system-level concern. Training large generative models and running extensive simulations consume energy and water, while automated synthesis and high-throughput assays generate chemical waste. Multi-objective optimization can include environmental impact as an explicit objective, encouraging routes that use greener solvents, fewer steps, and less hazardous reagents. Yet sustainability metrics are contested and data-dependent. Policy instruments such as carbon accounting, lifecycle assessment, and responsible research and innovation can help align molecular design with environmental goals. The same Bayesian machinery used to model affinity uncertainty can be used to model uncertainty in toxicity, environmental fate, or supply-chain risk, making trade-offs more explicit.

8. Case Illustrations and Cross-Domain Comparisons

Several case domains illustrate the promise and difficulty of multi-objective generative optimization. In kinase inhibitor discovery, deep learning has been used to identify potent compounds rapidly [4], but selectivity across the kinome remains a major challenge. A generative system must optimize affinity for the target while penalizing binding to closely related kinases. Simulation can model conserved and divergent pocket residues, while Bayesian active learning can select compounds that discriminate between targets. The system must also consider pharmacokinetics and safety, because a potent but non-selective inhibitor may fail clinically. This illustrates why affinity optimization cannot be separated from the broader decision architecture.

Antibiotic discovery provides another case. Deep learning enabled the discovery of a new antibacterial compound [3], but antibiotic development faces scientific, economic, and policy obstacles. Generative optimization could propose molecules that evade resistance mechanisms, but it must also consider spectrum, resistance propensity, and manufacturing feasibility. The system's value depends on validation pipelines and access policies. If a promising candidate is identified but cannot be produced or distributed equitably, the optimization system has not

solved the socio-technical problem. Similar issues arise in antiviral and antiparasitic discovery, where market incentives may be weak and public health needs are high.

Cross-domain comparisons with materials discovery and protein engineering reinforce these lessons. Inverse molecular design for materials has shown how generative models can propose candidates with desired properties [20], but materials optimization often tolerates longer development cycles and different safety constraints. Protein design has advanced through deep network hallucination and all-atom generative methods [7], but proteins and small molecules differ in synthesis, immunogenicity, and delivery. The common pattern is a closed loop that combines generative search, physical evaluation, experimental feedback, and multi-objective governance. The differences lie in the cost structures, regulatory pathways, and social values attached to each domain. A mature research program would develop reusable abstractions for these loops while remaining sensitive to domain-specific constraints.

9. Future Directions

Future progress will likely depend on foundation models that integrate sequence, structure, ligand, assay, and textual knowledge. Such models could support transfer learning across targets and tasks, but they require careful curation and evaluation. Bayesian flow and multi-modal generative methods offer promising routes for combining discrete and continuous representations while preserving uncertainty [7], [12]. Equivariant and graph-based architectures will remain important for physical consistency [8], [10]. Simulation will become more tightly coupled with generation through differentiable or surrogate models, but the epistemic status of simulation outputs must remain explicit. Benchmarks should measure not only predictive accuracy but also calibration, robustness, diversity, synthesis feasibility, and decision quality.

Deployment will require stronger governance. Model cards, datasheets, audit trails, and uncertainty reports can make systems more accountable. Regulatory agencies will need methods to evaluate generative design in drug submissions, including how candidates were selected and how uncertainty was handled. International coordination can address dual-use risks and equitable access. Education should train scientists to work across machine learning, simulation, and policy, because the hardest problems sit at interfaces. The most valuable future systems will not be those that generate the most molecules, but those that generate trustworthy knowledge and actionable candidates within a defensible decision process.

10. Conclusion

Multi-objective generative optimization of protein–ligand binding affinity using Bayesian inference and molecular simulation represents a convergence of computational statistics, physical modeling, artificial intelligence, and institutional design. Bayesian inference provides a language for uncertainty and evidence accumulation. Molecular simulation provides physical grounding and constraints. Generative models provide expansive search over chemical and conformational space. Multi-objective optimization provides a framework for navigating trade-offs among potency, selectivity, safety, synthesizability, and sustainability. None of these components is sufficient alone, and their integration creates new challenges in architecture, deployment, robustness, fairness, and policy.

The central conclusion is that affinity optimization should be treated as a socio-technical infrastructure. Its success depends on data provenance, simulation fidelity, uncertainty calibration, human oversight, laboratory validation, and governance. Systems that ignore these dimensions may produce impressive demonstrations but fragile or inequitable outcomes.

Systems that embrace them can support more reliable discovery, more transparent decision-making, and more equitable distribution of benefits. The path forward requires not only better models but also better institutions for evaluating, regulating, and sharing the results of generative molecular design.

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. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Denny, H., Peterson, S., Simonyan, K., Crossan, S., White, T., Vinyals, O., Kavukcuoglu, K., Hassabis, D., & Sifre, L. (2020). Improved protein structure prediction using potentials from deep learning. Nature, 577(7792), 706-710.

3. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. Cell, 180(4), 688-702.

4. Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev, A., Volkov, Y., Zholus, A., Shayakhmetov, R. R., Zhebrak, A., Minaeva, L. I., Zagribelnyy, B. A., Lee, L. H., Soll, R., Madge, D., ... Aspuru-Guzik, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nature Biotechnology, 37(9), 1038-1040.

5. Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., & Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. ACS Central Science, 4(2), 268-276.

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

7. 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.

8. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. Proceedings of the 34th International Conference on Machine Learning, 70, 1263-1272.

9. Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. Advances in Neural Information Processing Systems, 28, 2224-2232.

10. 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.

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

12. Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. International Conference on Learning Representations.

13. Rezende, D. J., Mohamed, S., & Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. Proceedings of the 31st International Conference on Machine Learning, 32, 1278-1286.

14. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623.

15. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

16. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

17. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery, 18(6), 463-477.

18. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug Discovery Today, 23(6), 1241-1250.

19. Schneider, P., Walters, W. P., & Plowright, A. T. (2020). Rethinking drug design in the artificial intelligence era. Nature Reviews Drug Discovery, 19(5), 353-364.

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