

Active Learning for Sample-Efficient Protein Design Using Bayesian Uncertainty Estimation and Experimental Feedback

Noemi Clerc

Department of Computer Science, University of North Texas, Denton, TX, USA.

noemiclerc@unt.edu

Keely McCadden

Department of Computer Science, George Mason University, Fairfax, VA, USA.

mccadden1993@gmu.edu

Tore Borgstrom

Department of Computer Science, University of New Hampshire, Durham, NH, USA.

tore.work@unh.edu

Abstract

Protein design is increasingly mediated by machine learning, yet experimental validation remains expensive, slow, and noisy. This paper examines active learning as a systems strategy for sample-efficient protein design, with particular attention to Bayesian uncertainty estimation and the integration of experimental feedback. Rather than treating uncertainty as a narrow modeling accessory, the discussion positions it as a control signal that governs which proteins are synthesized, assayed, and fed back into the next design cycle. The paper develops a system-level account of closed-loop protein design that spans assay orchestration, data provenance, model maintenance, acquisition policy, robotic execution, and governance. It analyzes key trade-offs between sequential and batched experimentation, computational and wet-lab costs, model calibration and exploration, and local optimization and long-term scientific value. The argument emphasizes that sample efficiency is not only a machine learning property but also an infrastructural, institutional, and policy outcome. Robustness under distribution shift, assay variability, and multi-objective design constraints is considered alongside fairness, biosafety, sustainability, and deployment in resource-constrained settings. Cross-domain comparisons with materials discovery, autonomous laboratories, and clinical machine learning clarify both the promise and the limits of Bayesian active learning. The paper concludes that effective protein design systems require uncertainty-aware models, interoperable data infrastructures, reproducible experimental workflows, and governance mechanisms that align efficiency with safety and equitable benefit. Such systems can reduce experimental waste and accelerate discovery, but only if uncertainty estimation, feedback quality, and institutional accountability are treated as first-class design concerns.

Keywords

active learning; protein design; Bayesian uncertainty; experimental feedback; sample efficiency; research infrastructure; governance; sustainability.

1. Introduction

Machine learning has transformed computational biology by enabling predictive and generative models that operate over complex biological representations [1], [2], [3]. Protein design is a particularly demanding application because the space of possible sequences, structures, and functions is vast, while experimental validation remains constrained by cost, time, and physical feasibility. Each wet-lab assay consumes reagents, instrument time, and expert attention, and many designed proteins fail to express, fold, or function as intended. In this setting, the central question is not whether models can propose candidates, but whether the overall socio-technical system can select candidates that maximize scientific progress per experiment. Active learning offers a disciplined answer by prioritizing experiments that are expected to reduce uncertainty or improve the design objective [13], [14], [15]. When coupled with Bayesian uncertainty estimation, active learning can distinguish between promising regions of design space and regions where the model is simply uninformed [8], [9], [10]. This distinction is crucial because protein fitness landscapes are often rugged, sparse, and noisy, and naive optimization can waste experimental capacity on misleading local peaks.

This paper argues that sample-efficient protein design should be understood as a systems problem rather than a purely algorithmic one. The model, the acquisition policy, the assay platform, the data infrastructure, and the governance regime form a single feedback loop. Weakness in any component can degrade the entire loop. A well-calibrated uncertainty model cannot compensate for inconsistent assays, and a high-throughput robotic platform cannot compensate for poor candidate selection. Similarly, an excellent acquisition policy may produce unfair or unsafe outcomes if deployment incentives and oversight mechanisms are misaligned [21]. The discussion therefore treats Bayesian uncertainty as a control signal that mediates between computational exploration and experimental commitment. It also examines how experimental feedback changes the model over time, how infrastructure shapes what can be learned, and how policy determines who benefits from accelerated design.

The paper contributes a long-form system-level analysis of active learning for protein design. It first reviews the problem setting and the role of uncertainty in protein engineering. It then describes a layered architecture for closed-loop design, including model services, experimental orchestration, and data provenance. Subsequent sections analyze Bayesian uncertainty as a control signal, experimental feedback and infrastructure, robustness and distribution shift, governance and fairness, sustainability and deployment, and evaluation across domains. The conclusion synthesizes these themes and proposes that sample efficiency must be evaluated alongside safety, reproducibility, and equitable access.

2. Background and Problem Setting

Protein design seeks to identify or create biomolecules with desired structural, catalytic, binding, or therapeutic properties. Traditional directed evolution iteratively mutates and selects variants, relying on assay throughput and selection pressure to navigate fitness landscapes. Machine-learning-guided directed evolution has improved this process by learning from limited experimental data and proposing variants that are more likely to succeed [6]. Structural models such as those used for protein structure prediction have further expanded the ability to reason about fold, interface, and function [2], [4]. Generative approaches can propose novel backbones and sequences, shifting the design problem from screening natural diversity to constructing candidates with specified properties [3]. These advances make the design space more tractable, but they also increase the number of plausible candidates, which raises the stakes for experimental prioritization.

The fitness landscape metaphor captures the difficulty of protein design without requiring formal mathematical treatment. A landscape may contain broad neutral regions, narrow peaks, epistatic interactions, and context-dependent effects. Experimental measurements are often noisy because expression, purification, assay conditions, and instrument calibration introduce variance. The same protein may perform differently across hosts, temperatures, or substrate concentrations. Consequently, a model trained on one batch may not generalize to the next. Active learning is valuable precisely because it accepts that data are limited and sequentially chooses the next experiment to maximize information or performance [13], [15]. Bayesian optimization provides a related framework, using a probabilistic surrogate model to balance exploration and exploitation [8], [9], [10]. In protein engineering, this balance determines whether the system searches broadly for unexpected solutions or refines variants around a known productive region.

Uncertainty estimation is central because protein design involves both model ignorance and experimental randomness. Epistemic uncertainty reflects what the model does not know because data are sparse, while aleatoric uncertainty reflects irreducible noise in the measurement or biological system. Methods such as dropout-based approximations and deep ensembles provide practical ways to estimate predictive uncertainty in deep learning [11], [12]. More recent work has explored multimodal Bayesian flow and all-atom representations to rationalize protein design under uncertainty [7]. Such approaches are important because they connect generative modeling with uncertainty-aware decision-making, but they do not by themselves solve the systems problem. The value of uncertainty depends on whether it is calibrated, whether it is used to select experiments, and whether the resulting feedback is captured in a form that improves future decisions. The remainder of this paper develops that systems perspective.

3. System Architecture for Closed-Loop Protein Design

A closed-loop protein design system can be understood as a layered architecture. The lowest layer includes experimental execution, such as DNA synthesis, expression, purification, binding assays, stability measurements, and functional screens. Above it sits an orchestration layer that schedules experiments, manages plates and instruments, and tracks sample identity. A data layer captures measurements, metadata, provenance, and quality control. A model layer hosts predictive, generative, and uncertainty models, while an acquisition layer converts model outputs into candidate rankings or batch designs. Finally, a governance layer defines safety checks, access controls, review thresholds, and audit trails. This architecture is not merely technical; it encodes institutional priorities about speed, reproducibility, safety, and cost. A system optimized only for throughput may underinvest in provenance, while a system optimized only for compliance may become too slow to support iterative learning.

The choice between sequential and batched experimentation is a defining architectural trade-off. Sequential active learning can adapt after each result, which is ideal when assays are fast and cheap. Batched active learning is necessary when synthesis or assay cycles take days or weeks, but batching reduces adaptability and can waste samples if the model is poorly calibrated. Robotic and autonomous laboratories offer a middle path by increasing throughput and reducing human error, as demonstrated by early robot scientist systems and subsequent automation platforms [18], [19]. However, automation also raises integration challenges. Instruments from different vendors may use incompatible data formats, and assay protocols may not be machine-readable. A successful architecture therefore requires interoperability standards, versioned protocols, and clear interfaces between computational proposals and

physical execution. Without these, the loop remains human-mediated and slow, limiting the practical benefit of sophisticated uncertainty estimation.

Model serving and training infrastructure also shape the loop. Transformer-based and related deep learning architectures have become common for sequence and structure representation [17], and optimization methods such as adaptive gradient descent enable efficient training at scale [16]. Yet deploying these models in a design loop requires more than training scripts. It requires reproducible environments, model versioning, monitoring for drift, and mechanisms for incorporating new experimental data without retraining from scratch. The system must also support human review, because not every proposed protein should be synthesized. Biosafety screening, ethical review, and intellectual property considerations may intervene before experimental execution. A mature architecture therefore treats the acquisition policy as a proposal engine rather than an autonomous decision-maker. It routes candidates through safety and feasibility filters, records the rationale for selection, and preserves enough context for later audit and learning.

4. Bayesian Uncertainty Estimation as a Control Signal

Bayesian uncertainty estimation provides a principled language for deciding when to trust a model and when to explore. In protein design, uncertainty can indicate that a candidate lies far from training data, that the model disagrees across ensemble members, or that the assay itself is noisy. These distinctions matter because the appropriate response differs. High epistemic uncertainty may justify an experiment that improves knowledge even if the predicted performance is modest. High aleatoric uncertainty may call for replicate measurements or improved assay protocols rather than additional design iterations. When uncertainty is used as a control signal, the acquisition policy can balance expected improvement, information gain, and risk. This is the core logic of Bayesian optimization and active learning [9], [10], [15]. The practical challenge is not the existence of acquisition functions but their calibration and integration with experimental realities.

Deep learning has made uncertainty estimation more scalable through approximate Bayesian methods and ensembles [11], [12]. These methods can produce predictive distributions over protein properties, but they can also be overconfident when training data are biased or when the model is misspecified. Calibration is therefore a first-class concern. A model that reports high confidence in a false prediction can cause the system to exploit a narrow region and miss better solutions. Conversely, a model that reports excessive uncertainty everywhere may cause the system to explore inefficiently. Recent work on multimodal Bayesian flow and all-atom protein design illustrates how uncertainty can be embedded in generative modeling rather than added after the fact [7]. This is a promising direction because it aligns candidate generation with uncertainty-aware selection. However, the system-level value of such methods depends on whether their uncertainties correspond to experimental outcomes and whether the experimental feedback is rich enough to update the model meaningfully.

Uncertainty also functions as a governance signal. If a proposed protein has high predicted functionality but also high uncertainty in safety-relevant properties, the system should route it to additional review or safer assays. If uncertainty is low but the candidate belongs to a dual-use category, policy constraints may still prohibit or restrict experimentation [21]. This means that Bayesian uncertainty cannot be the only control signal. It must be combined with rule-based safety checks, institutional review, and legal requirements. The most robust systems will treat uncertainty as one input among several, allowing different stakeholders to set thresholds for exploration, risk, and acceptable evidence. In this sense, uncertainty estimation

is not only a mathematical property of a model but also a boundary object that connects computational predictions, experimental decisions, and public accountability.

5. Experimental Feedback, Data Provenance, and Infrastructure

Experimental feedback is the lifeblood of active learning, but it is often the weakest link in protein design systems. Wet-lab data are heterogeneous, with different assays producing different scales, units, and reliability levels. Batch effects, reagent lot variation, and instrument drift can introduce systematic errors that models may mistakenly learn. Without metadata describing protocol versions, controls, and environmental conditions, feedback can be misleading. Data provenance is therefore essential. Each measurement should be linked to the sample, the protocol, the operator or robot, the instrument, and the software version used to process the result. This provenance supports reproducibility, error analysis, and continual learning. It also enables the system to weight or exclude unreliable data rather than treating all feedback as equally informative. In practice, provenance requires laboratory information management systems, electronic lab notebooks, and standardized ontologies that can travel across institutions.

Infrastructure choices shape what kinds of feedback are possible. Cloud-based model training offers elasticity and access to large computational resources, while on-premises systems may be necessary for sensitive data or low-latency control of instruments. Edge computing can support real-time assay analysis, but it may complicate model versioning and security. Autonomous laboratories integrate robotics, scheduling, and data analysis, reducing human intervention and enabling continuous operation [18], [19]. These systems are powerful but expensive and complex. They also concentrate capability in institutions that can afford them, raising concerns about equitable access [20], [22]. A more distributed model might combine centralized model services with local experimental nodes, allowing smaller laboratories to contribute data and receive improved designs. Such a model requires shared standards, privacy-preserving data exchange, and clear rules for attribution and benefit sharing.

Feedback quality also depends on experimental design. Replicates, controls, randomization, and blinding are standard in clinical and biological research, but they are not always prioritized in high-throughput protein engineering. Active learning can help by selecting experiments that discriminate between hypotheses, but it cannot compensate for fundamentally flawed assays. The system should therefore include quality control as part of the acquisition policy. For example, if the model is uncertain because of assay noise rather than biological novelty, the best next experiment may be a replication or calibration study rather than a new protein. This requires integrating statistical process control with active learning. It also requires human expertise, because assay development is often tacit and context-dependent. A system that ignores this expertise may optimize a proxy that is easy to measure but weakly related to the desired function.

6. Sample Efficiency, Robustness, and Distribution Shift

Sample efficiency is the primary promise of active learning, but it is not guaranteed. The degree of improvement depends on the quality of the surrogate model, the calibration of uncertainty, the diversity of the initial data, and the cost structure of experiments. In protein design, the cost of synthesis and assay may vary by orders of magnitude across candidates. A sample-efficient policy should therefore consider not only how many experiments are performed but also which experiments are worth performing. This is a multi-objective problem that includes predicted function, uncertainty, synthesis feasibility, toxicity, and

manufacturability. Bayesian optimization can handle multiple objectives in principle, but practical implementations often reduce them to a scalar score, which can hide important trade-offs [9], [10]. A system-level approach should make these trade-offs explicit and allow stakeholders to adjust them.

Robustness is another critical dimension. Models trained on one protein family may fail on another, and models trained under one assay condition may fail when conditions change. Distribution shift can arise from new hosts, new substrates, new environmental conditions, or new design objectives. Active learning can mitigate this by exploring regions where the model is uncertain, but it can also amplify bias if the acquisition policy over-explores a narrow region [14], [15]. Robustness requires monitoring for drift, maintaining diverse training data, and using ensembles or Bayesian approximations that reflect model disagreement [11], [12]. It also requires fallback strategies, such as reverting to directed evolution or human-designed variants when model predictions become unreliable. The system should be able to detect when it is out of distribution and adjust its behavior accordingly.

Distribution shift is not only a technical problem but also an institutional one. If different laboratories use different assays or standards, the combined dataset may be inconsistent. If data are not shared, each laboratory may remain in its own biased region. Federated learning and shared benchmarks can help, but they require governance and incentives. Researchers need reasons to contribute data and adhere to standards, such as access to improved models, authorship credit, or funding requirements. Without such incentives, the infrastructure fragments, and active learning systems cannot benefit from collective feedback. Cross-domain lessons from clinical machine learning and high-performance medicine are relevant here, because they show that data quality, validation, and equity cannot be separated from model performance [20], [22], [23]. Protein design systems should learn from these fields rather than repeating their mistakes.

7. Governance, Fairness, and Policy Implications

Governance is essential because protein design intersects with biosafety, biosecurity, public health, and intellectual property. The same methods that enable therapeutic discovery could be misused to create harmful biological agents. Active learning increases the risk surface by accelerating the search for functional proteins, including those with dangerous properties [5]. Policy responses should not stifle beneficial research, but they must include screening, access controls, and international coordination. Governance frameworks for artificial intelligence emphasize transparency, accountability, fairness, and human oversight [21]. These principles apply directly to closed-loop protein design. A system should document why candidates were selected, who approved them, what safety checks were performed, and how data were shared. It should also provide mechanisms for external review when decisions have high societal stakes.

Fairness is another concern. If advanced protein design systems are concentrated in wealthy institutions or countries, the benefits may bypass those with the greatest health needs. The data used to train models may come from populations or environments that are not representative, leading to designs that perform poorly in other contexts [22], [23]. Active learning can worsen this disparity if it prioritizes experiments that benefit only the institutions controlling the loop. A fairer approach would include diverse partners in problem definition, data collection, and benefit sharing. It would also support local capacity building, so that resource-constrained laboratories can use and contribute to these systems. Policy instruments

such as open standards, tiered access, and funding conditions can encourage this kind of distributed participation.

Governance must also address the tension between openness and security. Open science accelerates innovation and reproducibility, but some protein design data and methods may require controlled access. The challenge is to design tiered systems that allow broad participation while preventing misuse. This could include differential privacy for sensitive data, secure enclaves for dangerous sequences, and review boards for dual-use research. Active learning systems should be designed with these controls in mind, not retrofitted after deployment. For example, the acquisition policy could include safety classifiers that block certain candidates from synthesis, and the data layer could record all access and modifications. Such safeguards must be auditable and resistant to gaming. They also require international coordination, because biological research is global and regulatory frameworks vary across jurisdictions.

8. Sustainability and Deployment

Sustainability in protein design has both environmental and institutional dimensions. Wet-lab experiments consume reagents, plastics, energy, and water. Computational training also consumes energy, especially for large generative models. Active learning can improve sustainability by reducing the number of failed experiments, but it can also increase computational demand if models are retrained frequently or if acquisition policies are expensive to evaluate. A sustainable system should therefore measure the full lifecycle cost of both computation and experimentation. It should prefer candidate selection that reduces waste, and it should use efficient model architectures and hardware where possible [16], [17]. Cloud services can be efficient at scale, but their energy sources and cooling requirements vary. Institutions should consider these factors when deploying closed-loop systems.

Deployment in resource-constrained settings requires simplification and modularity. Not every laboratory needs a fully autonomous robotic platform. A modular system might begin with a predictive model, a simple acquisition policy, and manual experiments, then add automation as capacity grows. Open-source software, shared protocols, and low-cost instruments can lower barriers. However, sustainability also depends on maintenance. Models drift, instruments break, and personnel change. A system that cannot be maintained will not deliver long-term benefits. Deployment plans should therefore include training, documentation, spare parts, and local technical support. They should also include governance mechanisms that fit local contexts rather than imposing one-size-fits-all rules [21]. The goal is not merely to install technology but to build enduring capacity.

Sustainability also relates to scientific value. If active learning systems produce many publications but few deployable proteins or therapies, their societal return is limited. Evaluation should therefore include downstream impact, not just short-term metrics. This means tracking whether designed proteins are reproduced, whether they advance clinical or industrial applications, and whether benefits are shared. It also means considering the opportunity cost of focusing on narrow commercial targets rather than pressing global health needs. A sustainable protein design ecosystem would balance curiosity-driven exploration with mission-oriented problem solving. It would reward reproducibility, data sharing, and equitable collaboration as much as raw performance.

9. Evaluation and Cross-Domain Comparisons

Evaluating active learning for protein design requires metrics that reflect the full system. Common metrics include hit rate, number of experiments to reach a target, regret relative to the best known variant, and calibration error. These are useful but incomplete. A system should also be evaluated on reproducibility, time-to-result, cost per successful design, safety compliance, and distribution of benefits. Cross-domain comparisons can illuminate these dimensions. In materials discovery, Bayesian optimization has been used to navigate expensive experimental spaces [9], [10]. In autonomous laboratories, robot scientists have demonstrated closed-loop hypothesis generation and experimentation [18]. In synthetic biology, machine learning and automation have been combined to improve strain and pathway design [19]. These domains share the challenge of integrating noisy experiments, uncertain models, and physical constraints. They also share the risk of overfitting to easily measured proxies.

Clinical machine learning offers another comparison. High-performance medicine has shown that model accuracy alone does not guarantee clinical benefit; workflow integration, clinician trust, and equity matter greatly [20], [22], [23]. Protein design systems face similar socio-technical challenges. A model that recommends a protein with high predicted affinity may fail if the assay cannot measure that affinity reliably, if the protein cannot be manufactured, or if regulators require additional evidence. Evaluation should therefore include human factors and organizational readiness. It should also include failure analysis. Understanding why an experiment failed is often more valuable than recording that it failed. Active learning systems should be designed to capture negative results and use them to improve future decisions.

Cross-domain comparison also reveals differences. Protein design has a relatively mature set of structural and evolutionary priors, which can constrain the search space and improve sample efficiency. Materials discovery often lacks such priors, while clinical medicine has strong regulatory and ethical constraints. Protein design sits between these extremes: it is biologically grounded but still subject to dual-use concerns and regulatory oversight. A comparative perspective helps avoid overgeneralizing from one domain to another. The most robust evaluation frameworks will combine domain-specific assays with general principles of uncertainty, reproducibility, and governance. They will also include longitudinal studies that track whether active learning systems improve over multiple design cycles and whether their benefits persist after the original research team moves on.

10. Conclusion

Active learning for sample-efficient protein design is a promising approach that combines machine learning, Bayesian uncertainty estimation, and experimental feedback. Its success, however, depends on much more than the choice of acquisition function or model architecture. The entire system must be designed to select informative experiments, capture reliable data, update models appropriately, and route decisions through safety and governance checks. Bayesian uncertainty can serve as a control signal that balances exploration and exploitation, but it must be calibrated and integrated with experimental realities. Experimental feedback must be accompanied by provenance, quality control, and interoperable infrastructure. Deployment must consider sustainability, maintenance, and equitable access. Governance must address biosafety, biosecurity, fairness, and the tension between openness and control. This paper has argued that sample efficiency is a socio-technical achievement, not merely an algorithmic one. Systems that treat uncertainty, feedback, and governance as first-class concerns can reduce experimental waste, accelerate discovery, and distribute benefits more broadly. Systems that ignore these dimensions may optimize narrow metrics while creating

new risks and inequities. The next generation of protein design platforms should therefore be evaluated not only by how few experiments they require, but by how responsibly and reproducibly they convert those experiments into scientific and societal value.

References

1. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484-489. <https://doi.org/10.1038/nature16961>
2. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, 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. <https://doi.org/10.1038/s41586-021-03819-2>
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. <https://doi.org/10.1038/s41586-023-06415-8>
4. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Židek, A., Nelson, A. W. R., Bridgland, A., Simonyan, K., Vinyals, O., Wierstra, D., Silver, D., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792), 706-710. <https://doi.org/10.1038/s41586-019-1923-7>
5. Hie, B., Zhong, E. D., Berger, B., & Bryson, B. (2021). Learning the language of viral evolution and escape. *Science*, 371(6526), 284-288. <https://doi.org/10.1126/science.abd7331>
6. Yang, K. K., Wu, Z., & Arnold, F. H. (2019). Machine-learning-guided directed evolution for protein engineering. *Nature Methods*, 16(8), 687-694. <https://doi.org/10.1038/s41592-019-0496-6>
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. Romero, P. A., Krause, A., & Arnold, F. H. (2013). Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences*, 110(3), E193-E201. <https://doi.org/10.1073/pnas.1215251110>
9. Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 25, 2951-2959.
10. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1), 148-175. <https://doi.org/10.1109/JPROC.2015.2494218>

11. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *Proceedings of the 33rd International Conference on Machine Learning*, 48, 1050-1059.
12. 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.
13. Settles, B. (2009). Active learning literature survey. *Computer Sciences Technical Report 1648*. University of Wisconsin-Madison.
14. Angluin, D. (1988). Queries and concept learning. *Machine Learning*, 2(4), 319-342. <https://doi.org/10.1023/A:1022821128753>
15. Cohn, D. A., Ghahramani, Z., & Jordan, M. I. (1996). Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4, 129-145. <https://doi.org/10.1613/jair.295>
16. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations*.
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
18. King, R. D., Whelan, K. E., Jones, F. M., Reiser, P. G. K., Bryant, C. H., Muggleton, S. H., Kell, D. B., & Oliver, S. G. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971), 247-252. <https://doi.org/10.1038/nature02236>
19. Carbonell, P., Radivojevic, T., & García Martín, H. (2019). Opportunities at the intersection of synthetic biology, machine learning, and automation. *ACS Synthetic Biology*, 8(7), 1474-1477. <https://doi.org/10.1021/acssynbio.8b00540>
20. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44-56. <https://doi.org/10.1038/s41591-018-0300-7>
21. Floridi, L., & Cows, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>
22. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347-1358. <https://doi.org/10.1056/NEJMr1814259>
23. Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, 319(13), 1317-1318. <https://doi.org/10.1001/jama.2017.18391>