Protein Mutation Effect Prediction via Graph-Based Modeling of Residue pKa Shifts and Local Electrostatic Rewiring

Leon Satkings
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
leon843@missouri.edu

Lummy Lawrence
Department of Computer Science, University of North Texas, Denton, TX, USA.
contactlucas@unt.edu

Abstract

Predicting the functional and stability consequences of amino acid substitutions remains a cornerstone challenge in computational biology, with far-reaching implications for precision medicine, protein engineering, and the interpretation of genomic variation. While significant progress has been made by leveraging evolutionary sequence conservation and global structural features, the nuanced role of localized electrostatic perturbations, particularly those arising from residue-specific pKa shifts, has been comparatively underexplored in large-scale mutation effect prediction systems. This paper presents an interdisciplinary perspective on the design, deployment, and governance of graph-based modeling frameworks that explicitly encode pKa value alterations and the resultant electrostatic rewiring at protein interfaces. We articulate a system-level architecture in which proteins are represented as heterogeneous graphs with ionizable residues as charge-carrying nodes, and where edge attributes capture coulombic coupling and solvent exposure. By integrating physically inspired feature engineering with message-passing neural networks, such a system can capture the propagation of local charge disruptions across the contact network. The discussion extends beyond algorithmic design to address critical infrastructure demands, data provenance, robustness under distribution shift, fairness across protein families and human populations, model interpretability, and the policy frameworks required for clinical translation. We explore structural trade-offs between model granularity and computational tractability, the sustainability of training pipelines, and the ethical dimensions of embedding biophysical models into decision-support infrastructures. The analysis underscores that the value of such predictive platforms lies not solely in accuracy metrics but in the capacity to yield mechanistically transparent and socially accountable insights for variant interpretation at scale.

Keywords

protein mutation effect prediction, pKa shifts, electrostatic networks, graph neural networks, socio-technical systems, biomedical AI governance.

1. Introduction

The accurate prediction of how single amino acid substitutions modulate protein function, stability, and interaction propensity is a foundational problem in molecular biophysics with cascading consequences for clinical genomics, drug development, and synthetic biology. Tens

of millions of missense variants have been catalogued across human populations and model organisms, yet the overwhelming majority remain of uncertain significance, creating a vast interpretative bottleneck that no single experimental modality can resolve. Computational methods have thus become indispensable as screening tools, evolving from simple physicochemical scoring functions to sophisticated deep learning architectures that consume evolutionary profiles, three-dimensional structural coordinates, and biophysical simulation outputs. Among the determinants of mutation impact, the behavior of ionizable residues stands apart because their protonation states are exquisitely sensitive to the local microenvironment, and even modest perturbations to pKa values can reconfigure hydrogen bond networks, alter catalytic mechanisms, or drive pathological aggregation. Consequently, a predictive framework that places residue-specific pKa shifts and the attendant rewiring of local electrostatic contacts at its analytical center offers a route toward mechanistically richer variant effect prediction.

Historically, computational pKa prediction has been pursued through empirical methods such as PROPKA, which employ heuristic descriptors of desolvation and charge-charge interactions to estimate shifts in apparent pKa values relative to model compound references [3, 4]. These approaches, while rapid, operate under significant approximations that limit their transferability across protein families and conformational states. More recently, the emergence of deep generative models for protein sequences and structures has demonstrated that data-driven representations can implicitly capture complex epistatic and environmental dependencies that govern mutational landscapes [2, 1]. Graph neural networks, in particular, have proven to be a natural substrate for modeling proteins as relational systems wherein residues communicate through spatial proximity, hydrogen bonding, and non-covalent contacts. When node features are augmented with quantum-chemically informed or semi-empirical estimates of electrostatic potential, these architectures can, in principle, learn to predict mutation-induced pKa shifts without the explicit solution of the Poisson-Boltzmann equation on a per-variant basis. The challenge, however, is not confined to algorithmic innovation; it extends into the broader systems engineering of how such models are built, validated, maintained, and ethically embedded within translational pipelines.

This paper adopts a systems-oriented lens to examine the prediction of protein mutation effects via graph-based modeling of residue pKa shifts and local electrostatic rewiring. Rather than proposing a singular model, we interrogate the architectural principles, infrastructure requirements, robustness criteria, and governance imperatives that collectively determine whether such predictive systems can transition from academic benchmarks to societally beneficial tools. The following sections elaborate on the design space of graph architectures that fuse physical priors with learned representations, the data engineering ecosystems necessary to train and update these models, the multifaceted challenges of fairness and interpretability, and the policy frameworks that must co-evolve with technical capabilities.

2 .Related Work and System Context

Contemporary efforts to predict mutation effects span a spectrum from purely sequence-based approaches, which exploit the information contained in multiple sequence alignments, to structure-aware models that incorporate geometric and energetic descriptors. Deep generative models trained on family-level alignments have demonstrated remarkable capacity to score variant likelihoods in a manner that correlates with experimental fitness measurements, yet they remain agnostic to the physicochemical details of individual side-chain environments [2]. In parallel, graph-based architectures have been successfully deployed for tasks such as

protein design, model quality assessment, and binding site identification, demonstrating that message-passing operations over residue-level contact graphs can distill representations that reflect folding stability and interaction potential [1, 8]. However, the explicit integration of pKa dynamics and electrostatic network properties into these graph frameworks has been fragmented. Recent work has introduced graph-based deep learning models specifically designed to predict pKa values of protein-ionizable residues by engineering features that capture local backbone geometry, hydrogen bonding patterns, and solvent accessibility [6]. This development signals a maturation of the field toward the marriage of physical feature engineering and learned nonlinear composition, but its systematic coupling to mutation effect prediction remains an open frontier.

Macroscopic electrostatic models have long recognized that proteins organize their titratable groups into coupled networks where the protonation state of one residue can remotely influence another through both through-space and desolvation-mediated mechanisms [7]. When a mutation replaces or repositions an ionizable side chain, the effect is not merely a localized change in net charge; it propagates through the network, potentially flipping the protonation preferences of distant partners and altering the electrostatic landscape that governs ligand binding, allostery, and macromolecular assembly. Capturing such propagation within a graph model requires edge attributes that move beyond simple distance cutoffs to encode effective dielectric responses and orientation-dependent Coulombic interactions. This level of detail introduces substantial architectural trade-offs: richer edge representations improve fidelity to the underlying physics but dramatically expand the feature space, increasing the risk of overfitting and demanding larger, more diverse training corpora. The systems challenge lies in constructing a training regime and validation infrastructure that can discern whether improvements on held-out test sets reflect genuine physical insight or simply a model's capacity to memorize spurious correlations within the training distribution.

3 .System Architecture and Design Space

The envisioned system architecture centers on a heterogeneous graph representation of protein structures, wherein nodes represent both ionizable residues and their non-ionizable neighbors, and edges encode a spectrum of physical interaction types. Node features include residue identity, backbone dihedral angles, local electrostatic potential computed via a fast empirical method, and a set of solvation metrics derived from solvent accessible surface area calculations. Crucially, each ionizable residue carries an additional channel representing its reference pKa in aqueous solution and a predicted shift from that baseline, such that the model can condition mutation-effect predictions on the altered charge state probability at physiological pH. Edges between ionizable pairs are ascribed physically inspired attributes: the inverse of the distance-weighted dielectric, the dot product of unit vectors aligned with the side-chain dipole moments, and a binary flag indicating direct hydrogen bond participation. The message-passing layers are designed to permit the flow of charge-state information across the network, allowing the model to simulate, in a single forward pass, the global electrostatic rewiring that a point mutation would trigger.

Designing such an architecture forces careful consideration of the granularity at which pKa predictions are injected. A minimalist approach would treat pKa shifts as static pre-computed input features derived from an external tool, freezing the electrostatic component and limiting the model to learning higher-order combinatorial effects. A maximally expressive alternative would embed a differentiable pKa predictor as a sub-network, jointly training the full pipeline end-to-end, such that errors in pKa estimation are penalized by their impact on downstream

mutation effect accuracy. The latter strategy, while conceptually elegant, introduces significant computational overhead and optimization instability, particularly given the paucity of high-quality, experimentally determined pKa values across diverse protein microenvironments. A pragmatic middle ground involves pre-training a dedicated pKa prediction graph network on available spectroscopic and NMR datasets, then transferring its learned representations as frozen feature extractors into the mutation effect model, with occasional fine-tuning cycles triggered by the availability of new reference measurements. This modular decomposition mirrors best practices in large-scale machine learning infrastructure, where component reuse not only conserves computational resources but also facilitates independent validation and attribution of model behavior.

The architectural choices also intersect with questions of interpretability and uncertainty quantification. Graph attention mechanisms, which have been employed in protein quality assessment [8], offer a means to inspect which edges in the electrostatic network carry the greatest weight in determining a variant's predicted effect. By examining attention distributions in the final layers, researchers can generate hypotheses about critical charge-relay pathways that may be disrupted by clinically observed mutations. However, attention-based explanations in graph models are known to suffer from robustness issues, necessitating complementary methods such as GNNExplainer [17] to provide counterfactual perturbations of the input graph that would flip the predicted class. The system design must therefore incorporate a suite of post-hoc interpretability modules, each with well-defined scope and limitations, and expose their outputs through a user interface that does not overstate certainty.

4. Data Infrastructure and Engineering Considerations

The successful deployment of a mutation prediction system centered on electrostatic rewiring rests upon a robust data infrastructure that spans acquisition, curation, versioning, and continuous integration of heterogeneous biological data sources. The Protein Data Bank serves as the canonical repository for experimentally determined macromolecular structures, but its contents are biased toward well-studied proteins and conditions that may not reflect physiological solvent composition or crowding [10]. For each structure, extracting a complete representation of ionizable side chains requires that the deposited model exhibits sufficient resolution to resolve alternative conformations and ordered water molecules, criteria that exclude a large fraction of entries. Consequently, the infrastructure must integrate multiple tiers of structural data: high-resolution X-ray and cryo-EM structures for ground-truth electrostatic computations, homology models for protein families where experimental coverage is sparse, and the growing corpus of computationally predicted structures from initiatives such as AlphaFold [5]. The heterogeneity in coordinate accuracy across these sources propagates into pKa and electrostatic feature estimates, demanding explicit uncertainty calibration at the data preprocessing stage.

The construction of training labels presents its own engineering challenges. Experimental pKa values measured by NMR titration or spectrophotometric methods are available for only a limited subset of residues, predominantly in a few dozen model systems, and are aggregated in databases that vary in annotation standards. The scarcity of direct supervision forces the infrastructure to employ multi-task learning strategies, where the model simultaneously predicts mutation effects on thermodynamic stability, catalytic activity, and binding affinity using assay data that are far more abundant. This multi-objective setting raises subtle issues of gradient conflict and data imbalance that must be managed through adaptive loss weighting schedules and careful monitoring of per-task performance metrics during distributed training.

The computational demands of training graph networks on tens of thousands of structures with hundreds of residues each, combined with the overhead of on-the-fly feature recomputation during data augmentation, necessitate the use of GPU clusters with high-bandwidth interconnects and persistent caching of precomputed electrostatic maps. Decisions about whether to host the training infrastructure on institutional clusters, commercial cloud providers, or federated learning consortia carry significant implications for data governance, cost sustainability, and the ability to incorporate proprietary pharmaceutical datasets.

5. Robustness, Fairness, and Bias Mitigation

Any predictive system intended for use in variant interpretation must be scrutinized for its robustness under distribution shift and its potential to encode biases that could harm equitable clinical outcomes. Protein sequence space is deeply non-uniform: certain folds and families, such as kinases and immunoglobulins, are massively overrepresented in both structural databases and functional assay collections, while intrinsically disordered regions and membrane proteins remain underrepresented. A model trained predominantly on globular soluble domains may develop an implicit inductive bias that incorrectly penalizes charge-changing mutations in disordered linkers or transmembrane helices, where the dielectric environment and pKa reference states differ substantially. Such systematic errors could lead to the misclassification of benign polymorphisms as pathogenic in proteins that happen to fall outside the training distribution, with cascading consequences for patient diagnosis and genetic counseling.

Fairness concerns extend beyond protein biophysics to the human genetic context in which variant interpretation is applied. The reference human genome and the population databases from which allele frequencies are derived have historically skewed toward individuals of European ancestry, a disparity that is amplified in functional genomics resources [13]. If a mutation effect predictor incorporates allele frequency prior features or is calibrated against labeled variants from predominantly European cohorts, its performance may degrade when applied to individuals from underrepresented populations. Building a system that is robustly fair therefore requires deliberate architectural and data governance interventions: the inclusion of population-specific allele frequency tiers as conditional inputs, the active curation of diverse variant-labeling initiatives, and the enforcement of disaggregated performance reporting as a standard evaluation criterion. The computational cost of conducting such stratified evaluations across dozens of population groups, each with its own linkage disequilibrium pattern and variant spectrum, must be budgeted into the system’s continuous validation pipeline from its inception, rather than being relegated to a post-hoc audit.

Robustness to adversarial variation in protein structure inputs merits equal attention. Small perturbations in backbone coordinates, arising from differences in crystallization conditions or from the inherent uncertainty in deep learning-based structure prediction, can alter the calculated solvent accessibility and inter-residue distances that feed electrostatic feature computations. A system that produces sharply divergent pKa shift predictions for conformations within the expected error range of state-of-the-art predictors would be unreliable in the real-world setting where a single cryo-EM density map may support multiple plausible side-chain rotamer assignments. Techniques such as Monte Carlo dropout, deep ensembles, and test-time augmentation through local conformational sampling are necessary to convey predictive uncertainty to end users [12]. However, these techniques multiply inference latency, raising a tension between the clinical need for rapid variant screening and the scientific imperative for calibrated confidence. The resolution of this tension is as much a

matter of system engineering—through model distillation, caching of conformational ensembles, or asynchronous batch processing—as it is of algorithmic innovation.

6. Governance, Policy, and Ethical Translation

The translation of a research-grade mutation effect predictor into a component of clinical decision-making infrastructure activates a complex regulatory and ethical landscape that demands proactive governance design. In the United States, software systems that provide variant interpretation and are marketed for clinical diagnostic purposes may fall under the purview of the Food and Drug Administration, subjecting them to premarket review and postmarket surveillance requirements. Even when such tools are deployed within academic medical centers as laboratory-developed tests, their outputs increasingly influence patient management decisions, from the classification of variants for hereditary cancer syndromes to the nomination of actionable mutations in tumor sequencing. The governance framework must therefore specify the intended use population, the scope of variants and genes covered, and the mechanism by which updates—such as retraining on expanded data or correction of a biased feature—are validated and communicated to clinical users without disrupting ongoing care [18].

Transparency obligations extend to the explainability of predictions, a domain where graph-based electrostatic models offer a potential advantage. Because the models operate over structurally grounded interaction networks, they can produce mechanistic narratives that link a predicted deleterious effect to the disruption of a specific salt bridge or the deprotonation of a buried catalytic residue. However, rendering such narratives accessible to clinical geneticists, molecular pathologists, and genetic counselors requires careful interface design and the development of a shared lexicon that bridges biophysical computation and medical terminology. A failure to invest in human-centered design of explanation interfaces risks widening the gap between computational outputs and clinical utility, or, worse, engendering misplaced trust in mechanistic rationales that confound correlation with causation.

The sustainability and energy footprint of large-scale protein modeling also warrants policy consideration. Training state-of-the-art graph networks on extensive structural datasets, particularly when combined with ensemble methods for uncertainty estimation, can consume quantities of electricity and generate carbon emissions that conflict with institutional sustainability commitments [11]. While individual protein models remain modest in comparison to large language models, the broader trend toward exhaustive variant precomputation—evaluating every possible single amino acid substitution across the human proteome—carries a significant environmental toll if implemented naively. Governance frameworks must therefore encourage the sharing of precomputed model outputs through federated databases, the adoption of efficient inference architectures, and the transparent reporting of energy consumption alongside performance metrics. Such reporting aligns the biomedical AI community with the broader movement toward green artificial intelligence and ensures that the computational tools developed to improve human health do not inadvertently undermine planetary health.

A further layer of governance addresses the tension between proprietary model development, often fueled by pharmaceutical investment, and the open science norms that have historically underpinned structural biology and genomics. Models trained on public data but distributed as closed commercial services raise challenging questions about data sovereignty and the collective benefit of publicly funded resources. The structural bioinformatics community has increasingly advocated for model cards, datasheets, and open benchmarking platforms that

enable independent evaluation and prevent the lock-in of variant interpretation standards to a small number of commercial entities. Graph-based pKa shift models, because they rely on interpretable physical features rather than massive parameter counts, may be particularly amenable to open-source distribution and community-driven improvement, provided that the supporting data pipelines are designed with modularity and reproducibility from the outset [20].

7.Conclusion

The prediction of protein mutation effects through graph-based modeling of residue pKa shifts and local electrostatic rewiring represents a convergence of biophysical insight, graph machine learning, and large-scale data engineering that holds substantial promise for both mechanistic discovery and translational impact. This paper has argued that the realization of that promise depends on systems-level thinking that extends well beyond accuracy on curated benchmarks. The architectural choices governing how electrostatic information flows through graph networks, the infrastructure decisions that determine data quality and training reproducibility, the fairness and robustness criteria embedded in validation protocols, and the governance frameworks that guide clinical deployment are all integral components of a responsible predictive system. By designing models that treat electrostatic networks as first-class citizens rather than an afterthought, the field can move toward variant interpretation tools that are not only more accurate but also more interpretable and more just. The path forward requires sustained collaboration across disciplines, the establishment of shared open data resources that capture the full diversity of protein environments and human genetic variation, and a policy environment that incentivizes transparency and sustainability. In this vision, a graph-based electrostatic model is not merely a computational artifact; it is a node within a larger socio-technical network that, if carefully governed, can strengthen the connection between molecular data and human well-being.

References

1. Ingraham, J., Garg, V., Barzilay, R., & Jaakkola, T. (2019). Generative models for graph-based protein design. Advances in Neural Information Processing Systems, 32.
2. Riesselman, A. J., Ingraham, J. B., & Marks, D. S. (2018). Deep generative models of genetic variation capture the effects of mutations. Nature Methods, 15(10), 816–822.
3. Sondergaard, C. R., Olsson, M. H. M., Rostkowski, M., & Jensen, J. H. (2011). Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pKa values. Journal of Chemical Theory and Computation, 7(7), 2284–2295.
4. Bas, D. C., Rogers, D. M., & Jensen, J. H. (2008). Very fast prediction and rationalization of pKa values for protein-ligand complexes. Proteins: Structure, Function, and Bioinformatics, 73(3), 765–783.
5. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589.
6. Song, Z., Wang, R., Jiao, X., & Huang, Z. (2026). Graph-Based Deep Learning Models for Predicting p K a Values of Protein-Ionizable Residues via Physically Inspired Feature Engineering. Journal of Chemical Information and Modeling.

7. Bashford, D. (2004). Macroscopic electrostatic models for protonation states in proteins. Frontiers in Bioscience, 9, 1082–1099.

8. Baldassarre, F., Menéndez Hurtado, D., Elofsson, A., & Azizpour, H. (2021). GraphQA: protein model quality assessment using graph convolutional networks. Bioinformatics, 37(3), 360–366.

9. Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. Annual Review of Biomedical Data Science, 4, 123–144.

10. wwPDB consortium. (2019). Protein Data Bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Research, 47(D1), D520–D526.

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

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

13. Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. Nature, 538(7624), 161–164.

14. Schuhmacher, A., Gatto, A., Hinder, M., Kuss, M., & Gassmann, O. (2022). The state of artificial intelligence in biopharma 2022. Drug Discovery Today, 27(9), 2522–2529.

15. Kumar, S., & Nussinov, R. (2002). Close-range electrostatic interactions in proteins. ChemBioChem, 3(7), 604–617.

16. Reeb, J., Hecht, M., Mahlich, Y., Bromberg, Y., & Rost, B. (2020). Variant effect predictions incorporate both sequence and structure information. Bioinformatics, 36(12), 3637–3644.

17. Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. Advances in Neural Information Processing Systems, 32.

18. World Health Organization. (2021). Ethics and governance of artificial intelligence for health: WHO guidance. World Health Organization.

19. Karpov, P., Godin, G., & Tetko, I. V. (2020). Protein function prediction using graph neural networks and sequence embeddings. In 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 463–468). IEEE.

20. Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature, 533(7604), 452–454.