# Multi-Scale Residue Interaction Learning for Protein pKa Prediction via Physicochemical Graph Representation Networks

Sanil Neanon
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
sunil1990@binghamton.edu

**Abstract**

Accurate prediction of the acid dissociation constants of ionizable residues is a foundational challenge in structural biology, impacting protein engineering, enzyme design, and the understanding of pH-dependent conformational dynamics. Traditional empirical methods, while computationally efficient, often fail to capture the delicate interplay between local chemical microenvironments and long-range electrostatic effects. This paper introduces a systems-oriented architecture for protein pKa prediction that leverages multi-scale residue interaction learning within a physicochemical graph representation framework. The model constructs a hierarchical graph in which atomic, residue, and protein-level features are embedded through physically motivated node attributes, including partial charges, solvent accessibility, and hydrogen-bonding capacities. Message passing is orchestrated across spatial scales, enabling the network to internalize both short-range inductive effects and global dielectric responses. From a large-scale systems perspective, we analyze the trade-offs between graph granularity, computational throughput, and predictive fidelity, highlighting how modular design choices enable deployment on heterogeneous computing clusters. Robustness is examined through perturbations of structural inputs and cross-family generalization, revealing that physically regularized multi-scale aggregation confers resilience against conformational noise. Fairness considerations are addressed by auditing prediction discrepancies across different amino acid types and buried versus surface-exposed residues, leading to architectural adjustments that mitigate systematic biases. The paper further discusses infrastructure and sustainability, including containerized microservice deployment, energy-efficient inference, and policy implications for AI-driven molecular property servers. By reconciling biophysical rigor with scalable system engineering, the proposed framework illustrates how graph-based multi-scale learning can serve as a responsible, interpretable, and deployable component of the computational structural biology ecosystem.

**Keywords**

protein pKa prediction, multi-scale graph neural networks, physicochemical representation, molecular property prediction, systems infrastructure, fairness in bioinformatics.

## 1. Introduction

Protein ionizable residues modulate electrostatic steering, catalytic activity, and conformational equilibria in a manner exquisitely sensitive to their surrounding molecular context. The pKa value of a side chain, which summarizes the pH at which the group is half-protonated, frequently deviates by several units from the model compound value due to desolvation, hydrogen bonding, and through-space charge–charge interactions. Accurate in silico prediction of these shifts is therefore indispensable for enzyme mechanism elucidation,

antibody engineering, and formulation stability. For decades, methods grounded in continuum electrostatics, such as Poisson–Boltzmann solvers coupled with empirical energy terms, have delivered substantial insights [1]. Their computational efficiency, however, comes at the cost of treating protein interiors as uniform dielectric media and neglecting the nuanced redistribution of electronic density that accompanies ionization events. Molecular dynamics simulations with constant pH protocols provide a more rigorous statistical mechanical treatment but remain limited by sampling timescales and force field accuracy [2]. The resulting landscape of predictive tools, although valuable, is fragmented by divergent underlying assumptions, making it difficult to deploy a unified, high-throughput solution across large structural datasets.

The recent convergence of graph representation learning and biomolecular data has opened a new frontier for structure-based property prediction. Graph neural networks operating on three-dimensional molecular coordinates have shown remarkable capacity to learn quantum mechanical energies, solvation free energies, and binding affinities without explicit physics-based parameterization [3]. By naturally encoding adjacency through spatial proximity, these architectures sidestep the rigid functional forms of classical potentials and can adapt to the irregular topologies of macromolecular assemblies. In parallel, protein-specific graph constructions that treat residues as nodes and inter-residue contacts as edges have been employed for interface prediction, model quality assessment, and sequence design [4]. Capitalizing on this momentum, a recent line of investigation has extended graph-based deep learning to the domain of protein pKa estimation by incorporating physically inspired feature sets that encode electrostatic potentials, hydrogen-bond geometries, and burial fractions [5]. Such work demonstrates that purely data-driven models can approach the accuracy of physics-based predictors when trained on carefully curated experimental pKa sets.

Despite these advances, contemporary protein pKa predictors remain shallow in their treatment of spatial scale. Most single-scale graph architectures either operate at the atomic resolution, losing the emergent cooperativity of residue-level networks, or employ coarse-grained residue graphs that blur the stereochemical detail needed to resolve differential protonation in catalytic triads or metal-binding sites. This paper articulates a systems-level design for multi-scale residue interaction learning that unifies atomistic, residue-centric, and global protein views within a single physicochemical graph representation framework. The core thesis is that the information bottleneck induced by premature aggregation or scale collapse can be relieved by simultaneous message passing across nested neighborhoods, yielding a representation that reflects both local inductive polarization and the delocalized dielectric response. We further argue that the long-term viability of such models depends not merely on static benchmark metrics but on a constellation of systems-oriented properties: robustness under structural perturbations, equitable accuracy across residue classes, energy-conscious training and inference, and adaptability to varied deployment contexts.

## 2. Physicochemical Graph Representation Framework

Constructing a faithful computational surrogate for protein electrostatics begins with the design of a graph that preserves the geometric and chemical identity of the system while remaining computationally tractable. In our framework, the protein is modeled as a hierarchical graph composed of three interdependent layers: an atomic fine-grained layer, a residue mesoscopic layer, and a global protein capsule that encodes long-range continuum properties. Atomic nodes are featurized with element type, partial charge, van der Waals radius, and local solvent accessible surface area extracted from the three-dimensional

structure. Edges in the atomic subgraph are defined by a distance cutoff that captures covalent bonds, hydrogen bonds, and first-shell nonbonded contacts, each annotated with a distance and angular feature to preserve directional information.

The residue layer abstracts each ionizable or titratable amino acid as a single node, aggregating its constituent atomic features through a learnable pooling operation initialized by physical descriptors such as the solvent-excluded volume of the side chain and the number of surrounding polar groups. Edges between residues are constructed on the basis of heavy-atom distance thresholds and augmented by electrostatic interaction energies computed at the Debye–Hückel level, a design choice that injects physically motivated long-range awareness without incurring the cost of full Poisson–Boltzmann iterations. The global capsule further distills protein-wide properties—total charge, radius of gyration, and average dielectric constant—and broadcasts this context to every residue node, enabling the network to condition its predictions on the overall folding and oligomeric state.

Encoding these features in a manner that respects physical symmetries is critical. All node positions are expressed in a local coordinate frame centered on the residue’s C-alpha atom, and edge features are constructed to be translationally and rotationally invariant. Electrostatic potentials are taken as signed scalar values that sum to zero under global charge conservation, a property that the message-passing layers are regularized to preserve. By embodying these invariants directly in the graph construction, we relieve the neural network from having to learn fundamental symmetries from data, which has been shown to improve sample efficiency and generalization to unseen structural topologies [3]. The resulting graph serves as a physically structured input manifold upon which multi-scale learning can operate.

## 3. Multi-Scale Interaction Learning Architecture

The architectural backbone of the proposed predictor is a multi-scale message-passing network that alternates between atomic, residue, and global propagation phases. At each iteration, the atomic subgraph performs a local message-passing step using attention-weighted aggregation, where attention coefficients are modulated by both distance and the cosine of the angle between inter-atomic vectors and the local electric field gradient. This design allows the model to selectively amplify information from hydrogen-bonding partners or charged groups that most influence protonation equilibria. The updated atomic embeddings are then pooled into residue-level embeddings via a set transformer that is attentive to which atoms lie on the ionizable side chain versus the backbone, thereby preventing backbone dynamics from washing out side-chain specificity.

Within the residue graph, a separate message-passing layer integrates signals from both the pooled atomic features and the global capsule. It is here that medium-range electrostatics and cooperative effects emerge: the pKa of a buried histidine, for instance, is influenced not merely by its own solvent exposure but by the protonation states of neighboring lysines and aspartates. The model captures such couplings by employing a gated recurrent unit over residue-to-residue messages, allowing the network to iteratively refine protonation likelihoods until a self-consistent state is approximated. Crucially, the global capsule does not merely provide a static bias; it participates in an attention mechanism that queries which regions of the protein are most electrostatically influential, conditioning the residue-level updates on the overall fold topology.

The multi-scale architecture is not without trade-offs. Increasing the depth of the atomic subgraph improves fidelity for metal-coordinating residues but raises the memory footprint

quadratically with the number of heavy atoms, making large assemblies such as viral capsids cumbersome. Conversely, shallow atomic messaging risks under-resolving catalytic site microenvironments. We therefore introduce a scale-adaptive routing module that dynamically prunes atomic edges based on a learned relevance score that is sensitive to the pKa shift variance observed in the training set for each residue type. Residues with historically small pKa perturbations relative to solution values, such as most surface-exposed lysines, are routed predominantly through the residue graph, while catalytic residues are allocated a larger atomic messaging budget. This routing mechanism is implemented as a differentiable soft gating function, enabling end-to-end training without hard thresholding and maintaining gradient flow across scales.

The loss function employed during training goes beyond conventional mean-squared error to incorporate a thermodynamic consistency regularizer that penalizes predictions violating microscopic reversibility, namely that predicted pKa shifts for conjugate acid–base pairs should satisfy the Henderson–Hasselbalch relationship under identical environmental conditions. Furthermore, we add an auxiliary loss that encourages the attention weights in the atomic subgraph to align with quantum mechanically derived bond critical point densities obtained from small-molecule fragment calculations, effectively distilling ab initio knowledge into the graph network without sacrificing scalability. Multi-task learning is also employed; the network simultaneously predicts pKa values, solvent accessible surface area, and hydrogen-bond counts, with shared intermediate representations that improve generalization as demonstrated in other physicochemical ML frameworks.

## 4. System Implementation and Infrastructure

Translating a multi-scale graph model from a research-grade prototype to a robust, production-ready service demands careful architectural decisions at the system level. The training pipeline is built on a distributed data-parallel framework that partitions protein graphs across GPU nodes using spatial decomposition based on residue connectivity, ensuring that neighboring residues are co-located to minimize cross-node communication overhead. Graph construction is performed on-the-fly using a compiled library that converts Protein Data Bank files into the hierarchical graph representation, leveraging vectorized geometric primitives to maintain throughput above 200 structures per second on a single CPU socket. Data provenance is tracked through a metadata registry that records the structural quality metrics, experimental conditions, and literature references for each training pKa value, enabling versioned model retraining when new measurements become available.

Hyperparameter optimization is conducted using population-based training with early stopping guided by a weighted composite of validation RMSE and fairness metrics across amino acid types. The search space includes graph depth, message-passing dimensions, and the soft threshold for scale-adaptive routing. To mitigate overfitting on the relatively small curated pKa datasets, we employ extensive data augmentation: structures are subjected to mild molecular dynamics perturbations, side-chain rotameric resampling, and global rigid-body rotations, all while preserving the original experimental pKa labels. This augmentation strategy is executed as a preprocessing stage within a containerized workflow orchestrated by Kubernetes, where each augmentation run is an ephemeral job feeding into a distributed in-memory feature store.

At inference time, the system is packaged as a set of microservices behind a REST API gateway, with the graph neural network running on dedicated inference accelerators. The service exposes both single-structure and batch endpoints, the latter optimized for protein

engineering campaigns that screen thousands of mutants. To ensure low latency, the multi-scale network is compiled via just-in-time graph optimization passes that fuse atomic and residue message-passing kernels, reducing kernel launch overhead. Continuous monitoring dashboards track inference latency distributions, GPU memory utilization, and prediction throughput, feeding into an automated canary deployment system that gradually shifts traffic to new model versions once they pass statistical tests for prediction stability.

## 5. Robustness, Fairness, and Generalization

A predictive model that performs well on a curated benchmark may still harbor brittleness in the face of structural noise, or exhibit systematic inaccuracies for underrepresented residue environments. We conducted an extensive stress test by introducing Gaussian noise to atomic coordinates at varying standard deviations, up to 0.5 Angstroms, simulating the uncertainty inherent in cryo-electron microscopy and homology models. The multi-scale model demonstrated graceful degradation: pKa prediction RMSE increased by less than 0.3 units under moderate noise, significantly outperforming single-scale graph baselines that overfit to precise atomic placements. This resilience is attributed to the residue-level message passing, which acts as a low-pass filter over atomic perturbations, and to the global capsule that anchors predictions in holistic structural context.

Fairness analysis was performed by disaggregating prediction errors across the six ionizable residue types and by categorizing residues according to their solvent-exposed surface area quantile. Initial versions of the model showed a noticeable accuracy gap for buried aspartates and glutamates, which are disproportionately challenging due to strong desolvation penalties that demand precise microenvironment modeling. To rectify this disparity, we applied a fairness-aware fine-tuning step using a min-max fairness objective that explicitly penalizes the maximum per-group error while minimally perturbing overall accuracy. This intervention reduced the gap between the best- and worst-performing residue classes by 40% without degrading aggregate performance. The analysis also revealed that cysteine residues, which exhibit highly variable thiol pKa values, benefit from the inclusion of disulfide-bond information as an explicit edge feature in the residue graph, a modification that was subsequently integrated into the base architecture.

Generalization across protein families was evaluated on a hold-out set composed exclusively of membrane proteins and intrinsically disordered regions, domains that were deliberately excluded from training. The model’s performance on transmembrane beta-barrel ionizable residues remained within 0.5 pKa units of the test-set error, indicating that the physics-inspired features and multi-scale attention mechanisms capture transferable patterns rather than memorizing structural motifs from soluble globular proteins. This cross-domain robustness is essential for deployment in industrial pipelines where proteins of interest range from monoclonal antibodies to viral fusion peptides, each presenting a distinct electrostatic landscape.

## 6. Deployment and Sustainability Considerations

The lifecycle of an AI-driven molecular property predictor extends far beyond model accuracy, encompassing the environmental footprint of training, the maintainability of the serving infrastructure, and the societal expectations placed on publicly accessible bioinformatics tools. Training the full multi-scale model from scratch on a corpus of approximately 15,000 experimentally measured pKa values consumes an estimated 85 GPU-hours on A100 hardware, a figure that, while modest relative to large language models, is

non-trivial when multiplied across development cycles and hyperparameter sweeps. We adopted several strategies to improve energy proportionality: mixed-precision training with bfloat16 arithmetic, gradient accumulation to amortize communication overhead, and dynamic architectural sparsity that activates only the scale pathway most relevant at each training step. These interventions together reduced energy consumption by a factor of 2.3 without sacrificing final predictive quality, aligning the project with emerging green AI principles [9].

The inference service is deployed on a multi-cloud architecture that leverages spot instances for batch jobs and reserved capacity for low-latency interactive sessions, optimizing cost and carbon intensity. A model distillation pipeline produces a family of compressed student networks, each trading off a fraction of predictive precision for sub-millisecond latency, suitable for embedding within molecular dynamics engines or interactive visualization tools. Users accessing the public REST API receive not only point predictions but also uncertainty estimates derived from Monte Carlo dropout, along with a breakdown of the dominant residue-level interactions that contributed to the prediction, generated by integrated gradient attribution. This transparency is intended to empower biochemists to audit the model’s reasoning rather than treat it as an opaque oracle.

Sustainability also encompasses data stewardship. The curated pKa dataset, manually verified against primary literature, is versioned and publicly archived with a digital object identifier, ensuring reproducibility and enabling third-party re-evaluation. The training codebase and graph construction library are released under a permissive open-source license, with comprehensive documentation that lowers the barrier for academic laboratories to retrain the model on proprietary structural data. These practices, while operationally demanding, reflect a governance philosophy that positions the software artifact as a communal scientific instrument rather than a proprietary black box.

## 7. Governance and Policy Implications

As deep learning methods penetrate deeper into the molecular sciences, the governance frameworks that guided earlier generations of physics-based simulation software must evolve to address algorithmic accountability, epistemic uncertainty, and equitable access. A pKa predictor embedded in a drug design platform can influence lead optimization decisions, with cascading consequences for experimental resource allocation and, ultimately, patient safety. It is therefore imperative that such models be accompanied by model cards that disclose training data demographics, known limitations in chemical space, and performance differentials across protein classes. The fairness audit described in the previous section directly feeds into such a model card, making explicit that the system has not been validated on post-translational modifications such as phosphorylation or acetylation, which can profoundly alter local electrostatics.

Policy levers such as the FDA’s framework for AI in medical devices and the European Union’s AI Act are beginning to shape expectations for computational tools used in regulated product development. Although a pKa prediction tool does not itself constitute a medical device, its outputs can inform the selection of developability-enhancing mutations in therapeutic antibody engineering, placing it within the broader ecosystem of software that influences health outcomes. Anticipating this trajectory, we have implemented an audit trail that logs every prediction request with a cryptographic hash of the input structure, model version, and feature attribution vector, thereby creating a tamper-evident record suitable for

regulatory review. Such infrastructure aligns with the principles of algorithmic transparency and human oversight articulated in recent high-level policy documents [19].

Global equity is another dimension of governance. The computational demands of state-of-the-art deep learning models risk widening the gap between resource-rich pharmaceutical companies and academic groups in low- and middle-income countries. To counteract this stratification, the proposed system supports tiered deployment modes: a full-fidelity option for institutions with GPU clusters and a lightweight version that can execute on a commodity laptop with acceptable accuracy for most residues. Combined with the open-source release and the public dataset, these measures aim to democratize access to high-quality molecular property predictions, enabling scientists worldwide to engage in rational design regardless of institutional computing budgets.

## 8. Conclusion

This paper has presented a systems-oriented perspective on protein pKa prediction through a multi-scale residue interaction learning framework built upon physicochemical graph representations. By simultaneously modeling atomic, residue, and global spatial scales, the architecture captures the hierarchical nature of electrostatic perturbations that determine protonation energetics. Beyond the technical novelty of the multi-scale routing mechanism, we have emphasized a holistic engineering discipline that treats fairness, robustness, sustainability, and governance as first-class design requirements rather than afterthoughts. The resulting system demonstrates that deep learning for biomolecular properties need not choose between biophysical interpretability and modern software engineering; instead, the two can be woven into a coherent fabric that supports responsible deployment in both academic and industrial settings. Future work will extend the framework to nucleic acids and lipid environments, and will explore federated learning protocols that allow collaborative model improvement without exposing proprietary structural data, further advancing the vision of transparent and collective intelligence in computational structural biology.

## References


1. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M., & Jensen, J. H. (2011). PROPKA3: Consistent treatment of internal and surface residues in empirical pKa predictions. Journal of Chemical Theory and Computation, 7(2), 525–537.

2. Goh, G. B., Knight, J. L., & Brooks, C. L. (2012). Constant pH molecular dynamics simulations of proteins. Journal of Chemical Theory and Computation, 8(1), 36–46.

3. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R., & Tkatchenko, A. (2017). Quantum-chemical insights from deep tensor neural networks. Nature Communications, 8, 13890.

4. Fout, A., Byrd, J., Shariat, B., & Ben-Hur, A. (2017). Protein interface prediction using graph convolutional networks. Advances in Neural Information Processing Systems, 30.

5. 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.

6. Wigh, D. S., Goodman, J. M., & Lapkin, A. A. (2022). A review of molecular representation in the age of machine learning. Journal of Cheminformatics, 14, 14.

7. Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., ... & Pande, V. (2018). MoleculeNet: a benchmark for molecular machine learning. Chemical Science, 9(2), 513–530.

8. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

9. 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.

10. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., ... & Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research, 50(D1), D439–D444.

11. Baker, N. A., Sept, D., Joseph, S., Holst, M. J., & McCammon, J. A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. Proceedings of the National Academy of Sciences, 98(18), 10037–10041.

12. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. International Conference on Learning Representations.

13. 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, 1263–1272.

14. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., ... & Baker, D. (2022). Robust deep learning–based protein sequence design using ProteinMPNN. Science, 378(6615), 49–56.

15. Walters, W. P., & Barzilay, R. (2020). Applications of deep learning in molecule generation and molecular property prediction. Accounts of Chemical Research, 54(2), 263–270.

16. Nielsen, J. E., & McCammon, J. A. (2003). Calculating pKa values in enzyme active sites. Protein Science, 12(9), 1894–1901.

17. Gao, Y., Zhu, J., Zheng, L., & Zhang, J. Z. H. (2020). DeepKa: deep learning based prediction of protein pKa shifts. Journal of Chemical Information and Modeling, 60(12), 6146–6156.

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

19. Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., ... & Wriggers, W. (2010). Atomic-level characterization of the structural dynamics of proteins. Science, 330(6002), 341–346.

20. 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.