Structure-to-Function Learning: Predicting Enzyme Catalytic Residue Activity Through pKa-Aware Graph Representations

Rainer Burns
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
burnsrainer@buffalo.edu

Kannath Milkes
Department of Computer Science, University of North Texas, Denton, TX, USA.
kenneth.work@unt.edu

Nanshan Du
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
nanshandu@binghamton.edu

Abstract

The accurate prediction of enzyme catalytic residue activity from structural data stands as one of the central challenges in computational biology, with profound implications for drug discovery, industrial biocatalysis, and protein engineering. While recent advances in deep learning have enabled remarkable progress in protein function prediction, most existing methods either treat atomic-level chemical environments in an overly idealized manner or fail to integrate the critical physicochemical property of ionizable residue pKa values into their representations. This paper presents a framework for structure-to-function learning that constructs pKa-aware graph representations of enzyme active sites, jointly capturing three-dimensional spatial organization, evolutionary sequence features, and protonation-state energetics. At the system level, we examine the architectural trade-offs involved in encoding pKa information through physically inspired feature engineering within message-passing neural networks, including the balance between precomputed pKa predictions and on-the-fly electrostatic calculations. We provide a comprehensive analysis of the infrastructure required for large-scale deployment, spanning distributed training strategies, data provenance pipelines, and model-serving latency constraints in high-throughput screening contexts. Robustness is addressed through systematic evaluation of model sensitivity to structural perturbations, conformational sampling, and pKa prediction errors propagated from upstream modules. Fairness and bias considerations are discussed with respect to the overrepresentation of certain protein families in structural databases and the implications for generalizability to understudied enzyme classes. Sustainability concerns related to the computational footprint of training graph networks on massive structural datasets are evaluated alongside emerging efficient architectures. Finally, we outline governance and policy recommendations for the responsible dissemination of predictive models that could inform biocatalyst design, highlighting intellectual property boundaries, dual-use considerations, and the need for open benchmarking standards. Through this multidisciplinary lens, the work positions pKa-aware graph learning not merely as a technical advance in bioinformatics, but as a complex socio-technical system whose design choices reverberate through scientific practice, industrial deployment, and regulatory frameworks.

Keywords

enzyme function prediction, graph neural networks, pKa prediction, catalytic residues, structural bioinformatics, system design, computational sustainability, fairness in AI.

1. Introduction

Understanding the relationship between protein structure and catalytic function constitutes a foundational pursuit in molecular biology and a practical necessity for engineering enzymes with tailored properties. The catalytic activity of an enzyme is governed by a small set of residues whose chemical reactivity is exquisitely tuned by their local microenvironment, including hydrogen bonding networks, solvent accessibility, and electrostatic interactions. Among the most decisive yet often overlooked physicochemical parameters is the pKa value of ionizable side chains, which determines their protonation state under physiological conditions and thereby directly shapes nucleophilicity, electrophilicity, and acid-base catalysis mechanisms. Traditional computational approaches have treated catalytic residue identification as a sequence conservation or structural template matching problem, but the advent of deep learning, particularly graph neural networks that operate natively on protein topologies, opens the possibility of learning function directly from three-dimensional structure in a data-driven yet physically informed manner. This paper explores a framework termed structure-to-function learning, in which enzyme catalytic residue activity is predicted through graph representations that explicitly incorporate pKa values as node features or edge attributes, bridging the gap between quantum-level energetics and mesoscale structural learning.

The significance of this endeavor extends well beyond incremental prediction improvements. As the bioeconomy increasingly relies on engineered enzymes for sustainable chemical synthesis, bioremediation, and therapeutic applications, the ability to computationally screen vast sequence and structure spaces for catalytic competence becomes a critical infrastructure challenge. A pKa-aware graph learning system must be assessed not only by its accuracy metrics, but also by its scalability to the continuously expanding protein data universe, its robustness in the face of structural noise and conformational dynamics, and its fairness across the unevenly sampled phylogenetic tree. Furthermore, the computational resources required to train and operate such models raise important sustainability questions that demand a holistic systems perspective. This paper thus adopts the stance of an interdisciplinary systems researcher, examining the architectural decisions, data governance policies, deployment scenarios, and societal implications of pKa-aware graph representations for enzyme function prediction.

2. Background and Conceptual Foundations

The task of enzyme catalytic residue prediction has traditionally been addressed through sequence-based conservation analysis, where residues invariant across homologous families are inferred to be functionally important [1]. Structural bioinformatics added a powerful dimension by recognizing that catalytic residues often reside in particular spatial motifs, such as the catalytic triad in serine proteases, and exhibit characteristic geometries relative to substrates or cofactors [2]. Machine learning methods expanded the feature space to include solvent accessibility, secondary structure, and evolutionary profiles, using support vector machines or random forests to classify residues as catalytic or non-catalytic [3]. However, these approaches relied on handcrafted features that could not fully capture the complex, cooperative nature of the catalytic microenvironment.

The deep learning revolution in protein science, catalyzed by breakthroughs in structure prediction and representation learning, has shifted the paradigm toward end-to-end differentiable models that learn feature hierarchies directly from data [4]. Convolutional neural networks applied to three-dimensional density maps or distance matrices have been used for function prediction, but graph neural networks offer a particularly natural representation for proteins, which can be modeled as graphs where nodes correspond to amino acid residues and edges encode spatial proximity, covalent bonds, or interaction energies [5]. Message-passing operations on such graphs allow information to propagate across the structure, enabling the model to aggregate contextual cues from the entire neighborhood surrounding a candidate catalytic residue.

A crucial dimension often underrepresented in these graph models is the protonation chemistry of ionizable groups. The pKa of a residue such as histidine, cysteine, or aspartate can shift by several units from its solution value due to the protein environment, directly controlling whether the side chain is protonated and thus chemically active [6]. Computational methods for pKa prediction, ranging from Poisson-Boltzmann continuum electrostatics to empirical scoring functions, have been developed over decades, but their accuracy has been limited and their integration into deep learning pipelines remains unsystematic [7]. The emergence of graph-based deep learning models specifically designed for pKa prediction, which leverage physically inspired feature engineering to capture local electrostatic environments, marks a turning point that enables seamless incorporation of protonation state information into broader structure-to-function frameworks [15]. By embedding predicted pKa values as continuous features within the residue-level graph representation, a learning system can condition its function predictions on the protonation energetics that ultimately govern catalytic chemistry.

3. pKa-Aware Graph Representation of Enzyme Active Sites

Constructing a pKa-aware graph representation begins with the structural model of the enzyme, which may be derived from X-ray crystallography, cryo-electron microscopy, or computational prediction. Each residue is treated as a node, with initial features encoding amino acid identity, secondary structure class, solvent accessible surface area, and evolutionary conservation scores from multiple sequence alignments [8]. The critical extension is the inclusion of predicted pKa values for all ionizable residues, obtained from a pretrained graph neural pKa predictor that takes the local structural environment as input and estimates the microscopic pKa for each titratable side chain [15]. These pKa values serve as continuous physicochemical descriptors that inform the model about the likely protonation state under the assay conditions, effectively providing a window into the local electrostatic potential without requiring explicit quantum mechanical calculations.

Edges in this graph are constructed based on spatial distance cutoffs, typically linking residues whose C-alpha atoms or side-chain centroids fall within a threshold of 8 to 12 angstroms, thereby capturing both short-range hydrogen bonding and longer-range electrostatic effects [9]. Edge features may encode inter-residue distances, relative orientations of side chains, and the presence of non-covalent interactions such as salt bridges or pi-stacking. In the pKa-aware formulation, additional edge attributes can be derived from the difference in predicted pKa values between connected residues, which provides a coarse proxy for the proton transfer potential or for the likelihood of forming a charge-relay system characteristic of many enzyme mechanisms.

The graph construction process raises important system-level considerations regarding the source and quality of the pKa annotations. Embedding a pKa prediction model as a preprocessing stage introduces a dependency on its accuracy and generalization capacity; errors in pKa estimation can propagate and confound the downstream catalytic activity predictor. Therefore, the architecture must contend with uncertainty quantification and calibration, potentially through ensembles of pKa predictors or through training regimes that expose the downstream model to synthetic noise reflecting realistic pKa prediction errors [10]. This coupling between subsystems exemplifies the broader challenge of composing machine learning modules into robust analytical pipelines in computational biology.

4. Learning Framework and Architectural Design

The core learning module consists of a message-passing graph neural network that iteratively updates residue representations by aggregating information from neighboring nodes and edges. Multiple architectures can be instantiated, including graph convolutional networks, graph attention networks, and equivariant networks that respect three-dimensional rotational symmetries [5]. The choice of architecture introduces trade-offs between expressivity, computational cost, and interpretability. Attention-based mechanisms, for example, allow the model to learn which neighboring residues exert the greatest influence on a candidate catalytic residue's activity, yielding attention weights that can be inspected post hoc to generate mechanistic hypotheses. However, attention mechanisms increase memory consumption and may be less scalable to very large enzyme complexes or multi-chain assemblies.

The output head of the network is tailored to the specific prediction task, which may involve classifying each residue as catalytically active or inactive, regressing a continuous activity score, or forecasting the effect of mutations on catalytic efficiency. A crucial design decision concerns the level of supervision: fully supervised models require large, curated datasets of experimentally annotated catalytic residues, which are available from resources such as the Catalytic Site Atlas and Mechanism and Catalytic Site Atlas databases [2]. Semi-supervised and self-supervised pretraining strategies, in which the graph encoder is first trained on a vast corpus of unlabeled protein structures to reconstruct masked residue properties or predict spatial adjacencies, have shown promise in leveraging the wealth of structural data while mitigating annotation scarcity [11]. These pretrained representations can then be fine-tuned on the catalytic residue prediction task, a paradigm that aligns with the transfer learning successes observed in natural language processing and computer vision.

Integrating pKa awareness into this pipeline requires careful handling of the prediction target's conditionality on pH. Enzyme assays are typically performed at a specific pH, and catalytic activity is inherently pH-dependent. A comprehensive system should therefore accept pH as an auxiliary input, modulating the pKa-derived protonation probabilities through a Henderson-Hasselbalch transformation. While we avoid explicit mathematical formulations, the conceptual implication is that the graph network must learn to interpret pKa values in the context of the environmental pH, effectively learning a nonlinear mapping from (pKa, pH) pairs to functional outcomes. This dual-input design adds a dimension of complexity to the training data requirements, as the model benefits from training examples spanning a range of pH conditions, which are not always uniformly available in public biochemical databases.

5. Data Infrastructure and Evaluation Strategy

The construction of training, validation, and test datasets for pKa-aware catalytic residue prediction demands rigorous curation protocols to avoid information leakage and to ensure representativeness. A common pitfall is the inclusion of homologous sequences or structures that share high similarity, which can inflate performance estimates if not properly partitioned by sequence or structural clusters [12]. We advocate a clustering-based split strategy that groups proteins by sequence identity or CATH superfamily, ensuring that test sets contain enzymes evolutionarily and structurally distinct from those seen during training. This approach provides a more realistic measure of generalization to novel enzyme families, which is ultimately the use case of interest for enzyme discovery and engineering.

Evaluation metrics must extend beyond simple accuracy or area under the receiver operating characteristic curve, especially given the extreme class imbalance inherent in catalytic residue prediction, where the vast majority of residues are non-catalytic. Precision-recall curves, Matthews correlation coefficients, and per-family performance breakdowns are essential for diagnosing model behavior. A critical evaluation dimension is the robustness of predictions under structural perturbations, including those arising from conformational changes, mutations, or experimental resolution limitations. This can be probed by applying small random displacements to atomic coordinates or by sampling from molecular dynamics trajectories and observing the stability of predicted catalytic residue scores. A pKa-aware model should exhibit controlled sensitivity to structural perturbations that alter electrostatic environments, such as side-chain reorientations that break salt bridges, while remaining stable to inconsequential thermal fluctuations.

Reproducibility and transparency are cornerstones of the evaluation framework. All data preprocessing scripts, graph construction parameters, model architectures, and trained weights should be deposited in public repositories, with clear documentation of software dependencies and computational environments. The use of containerization technologies and model cards that detail performance characteristics, limitations, and intended use cases is recommended to facilitate independent verification and responsible downstream application [13].

6. System-Level Trade-offs: Scalability, Robustness, and Deployment

Deploying a pKa-aware graph learning system in a production environment, such as an industrial enzyme engineering platform or a public web server for academic users, introduces a distinct set of system-level challenges that transcend model development. The computational cost of inference must be balanced against the required throughput and latency. While graph neural network inference for a single enzyme is relatively fast, high-throughput screening campaigns may involve evaluating millions of protein variants, each requiring graph construction, pKa prediction, and forward passes through the activity model. The preprocessing step of computing predicted pKa values for all ionizable residues constitutes a significant computational bottleneck, particularly if the pKa predictor itself is a deep network that operates on fine-grained atomic graphs. Strategies to amortize this cost include precomputing pKa values for static structural databases, caching results for frequently queried scaffolds, and designing lightweight pKa approximators that trade a small amount of accuracy for substantial speed gains [14].

Robustness considerations extend to the handling of incomplete or low-quality structural inputs. In many realistic scenarios, an experimental structure may not be available, and the user must rely on computationally predicted models. The pKa-aware graph model must be resilient to the errors inherent in predicted structures, a property that can be enhanced through training on a mix of experimental and predicted structures with appropriate data augmentation

[4]. Furthermore, the system should gracefully degrade when encountering post-translational modifications, cofactors, or non-standard residues that were sparsely represented in the training data, by emitting calibrated uncertainty estimates rather than overconfident mispredictions. Uncertainty quantification techniques, including Monte Carlo dropout, deep ensembles, and conformal prediction, should be integrated into the serving infrastructure to enable risk-aware decision-making in downstream applications [16].

The architecture of the serving system itself requires careful engineering. A cloud-based microservice architecture can provide elastic scaling to handle variable demand, with model inference separated from the web frontend and data storage layers. Alternatively, edge deployment of smaller distilled models may be appropriate for integration into laboratory software that operates without reliable internet connectivity. In either case, rigorous monitoring of model performance drift over time, as new data accumulates and the deployed model ages, is essential. The implementation of continuous integration and continuous deployment pipelines that periodically retrain the model on expanded datasets and automatically evaluate against held-out benchmarks can ensure that the system remains state-of-the-art and safe for ongoing use [17].

7. Governance, Fairness, and Policy Considerations

The predictive models discussed here are not merely computational artifacts; they are components of a broader socio-technical system that shapes scientific knowledge production and technological innovation. One critical dimension is fairness and bias: the protein structure databases that underpin model training are heavily skewed toward well-studied organisms, model species, and biomedically relevant enzyme classes, while vast swaths of microbial and extremophilic enzyme diversity remain structurally uncharacterized [18]. A pKa-aware graph model trained predominantly on mesophilic, eukaryotic enzymes may exhibit systematically degraded performance on thermophilic, psychrophilic, or halophilic enzymes whose catalytic machinery operates under distinct physicochemical regimes. Such performance disparities could reinforce existing biases in enzyme research and redirect funding and attention away from understudied organisms with unique biocatalytic potential. Mitigating these biases requires intentional data collection efforts to fill phylogenetic gaps, as well as algorithmic interventions such as domain adaptation and fairness-aware training objectives.

Data governance and intellectual property policies also demand careful attention. Many high-quality enzyme structural datasets are generated through publicly funded research and deposited in open repositories, yet the predictive models derived from them can be commercialized by private entities, raising questions about the equitable distribution of benefits from public data. The pKa prediction model itself, if trained on proprietary experimental pKa datasets, may be subject to restrictive licensing that hinders its incorporation into open-source enzyme function prediction pipelines [19]. Clear data use agreements, open licensing standards, and community-driven benchmarking initiatives are essential to foster an innovation ecosystem that balances commercial incentives with the scientific commons.

Dual-use concerns, while less acute than in some other domains of artificial intelligence, are not entirely absent. Advanced enzyme function prediction tools can be used to engineer biocatalysts for the synthesis of valuable chemicals, but the same capabilities could potentially be directed toward the design of enzymes that degrade materials or produce harmful compounds. The scientific community should engage in proactive discussion about the appropriate boundaries for model access, perhaps through tiered release strategies that

provide full capabilities to verified academic researchers while offering more restricted interfaces to unauthenticated users. Such governance mechanisms must be developed transparently and with international coordination to avoid fragmentation or regulatory arbitrage.

8. Sustainability and Future Trajectories

The environmental sustainability of large-scale deep learning in computational biology is an increasingly pressing concern. Training state-of-the-art graph neural networks on millions of protein structures consumes substantial amounts of electrical energy and contributes to carbon emissions, particularly when hyperparameter sweeps and ensemble methods are employed. The pKa prediction preprocessing step multiplies this footprint, as each structure must be processed through an additional deep model before entering the activity prediction pipeline. Although individual inference operations have negligible cost, the aggregate impact of high-throughput screening campaigns can become significant. Researchers and platform operators should transparently report the energy consumption and estimated carbon emissions of their training and inference workloads, adopting best practices from the broader machine learning community on efficiency and hardware selection [20].

Looking forward, several promising directions can enhance both the performance and system-level properties of pKa-aware enzyme function prediction. The integration of language-model-based protein representations with structural graph networks offers the potential to fuse evolutionary sequence information with three-dimensional geometry in a more seamless and transferable manner [11]. Advances in equivariant neural networks that respect physical symmetries can improve sample efficiency and robustness to coordinate frame choices. On the pKa front, the development of differentiable pKa prediction modules that can be jointly optimized with the downstream activity predictor would allow end-to-end learning and more coherent uncertainty propagation. The emergence of large-scale experimental datasets providing systematic measurements of catalytic residue activities across diverse enzyme families will be invaluable for grounding these computational advances in empirical reality.

The trajectory toward more autonomous, closed-loop enzyme engineering platforms, where predictive models guide mutagenesis, expression, and characterization in iterative cycles, raises the stakes for all the system-level attributes discussed here. Reliability, interpretability, and fairness will no longer be abstract virtues but hard requirements for systems that make consequential decisions about which experiments to perform and which enzyme variants to pursue. The interdisciplinary framing of pKa-aware graph learning presented in this paper aims to equip researchers, engineers, and policymakers with the conceptual vocabulary to navigate these intertwined technical and societal dimensions.

9. Conclusion

This paper has articulated a comprehensive vision for structure-to-function learning of enzyme catalytic residue activity through pKa-aware graph representations, situating the technical challenge within a broader systems and societal context. We have examined how the explicit encoding of protonation energetics derived from graph-based pKa predictors enriches the feature space available to graph neural networks, enabling more physiochemically grounded function prediction. The architectural, infrastructural, and evaluation considerations extend far beyond model accuracy, encompassing the robustness to structural noise, the fairness across enzyme families, the computational sustainability of large-scale deployment, and the governance frameworks that should guide the responsible dissemination and use of

such technologies. By treating the pKa-aware graph learning system as a complex socio-technical infrastructure, we have highlighted the multidimensional trade-offs that must be navigated to realize the transformative potential of artificial intelligence in enzyme science while safeguarding scientific equity, environmental responsibility, and public trust.

References

1. Punta, M., Rost, B., & Ofran, Y. (2012). The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. PLoS Computational Biology, 8(10), e1002733.

2. Furnham, N., Holliday, G. L., de Beer, T. A. P., Jacobsen, J. O. B., Pearson, W. R., & Thornton, J. M. (2014). The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. Nucleic Acids Research, 42(D1), D485–D491.

3. Cilia, E., & Passerini, A. (2010). Automatic prediction of catalytic residues by modeling residue structural neighborhood. BMC Bioinformatics, 11, 115.

4. 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.

5. Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., ... & Bonneau, R. (2021). Structure-based protein function prediction using graph convolutional networks. Nature Communications, 12(1), 3168.

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

7. Anandakrishnan, R., Aguilar, B., & Onufriev, A. V. (2012). H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. Nucleic Acids Research, 40(W1), W537–W541.

8. Torng, W., & Altman, R. B. (2019). 3D deep convolutional neural networks for amino acid environment similarity analysis. BMC Bioinformatics, 20(1), 287.

9. Hermosilla, P., Schäfer, M., Lang, M., Fackelmann, G., Vázquez-Reina, A., Kozlíková, B., ... & Ropinski, T. (2021). Intrinsic-extrinsic convolution and pooling for learning on 3D protein structures. In International Conference on Learning Representations.

10. Slupsky, J. D., & Derewenda, Z. S. (2017). Machine learning approaches to protein pKa prediction. Current Opinion in Structural Biology, 43, 131–137.

11. Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., ... & Song, Y. S. (2019). Evaluating protein transfer learning with TAPE. In Advances in Neural Information Processing Systems.

12. Walsh, I., Pollastri, G., & Tosatto, S. C. E. (2014). Correct machine learning on protein sequences: a peer-reviewing perspective. Briefings in Bioinformatics, 15(5), 817–826.

13. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vassilev, L., Ozkaya, E., ... & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220–229).

14. Chen, T., Guestrin, C., & Rojas, F. (2022). Accelerating biomolecular deep learning with lightweight surrogate models. Nature Computational Science, 2(8), 521–529.

15. 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.

16. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In International Conference on Machine Learning (pp. 1050–1059).

17. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Young, M. (2015). Hidden technical debt in machine learning systems. In Advances in Neural Information Processing Systems.

18. Leman, J. K., Weitzner, B. D., Lewis, S. M., Adolf-Bryfogle, J., Alam, N., Alford, R. F., ... & Bonneau, R. (2020). Macromolecular modeling and design in Rosetta: recent methods and frameworks. Nature Methods, 17(7), 665–680.

19. Scudellari, M. (2021). Big data's big bias problem. Nature, 595(7866), S6–S8.

20. Strubell, E., Ganesh, A., & McCallum, A. (2020). Energy and policy considerations for modern deep learning research. In Proceedings of the AAAI Conference on Artificial Intelligence (pp. 13693–13696).