Multimodal Fusion of Sequence, Structure, and Electrostatic Features for Protein Ionization State Modeling

Mikkel Graves
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
mikkelg@unr.edu

Arthur Lindberg
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
lindbergarthur@oregonstate.edu

Lingtian Jia
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
lingjia80@unh.edu

Abstract

Accurate modeling of protein ionization states is a foundational challenge in computational biophysics, with direct implications for understanding enzyme catalysis, pH-dependent protein stability, and rational drug design. Traditional physics-based methods, while grounded in rigorous continuum electrostatics, often struggle with accuracy and scalability when applied across the growing corpus of protein structures. Recent advances in deep learning offer new pathways, yet the heterogeneity of relevant data sources—amino acid sequence, three-dimensional structure, and local electrostatic environments—demands integrative architectures capable of true multimodal reasoning. This paper presents a systems-level analysis of multimodal fusion frameworks for protein ionization state prediction, emphasizing architectural trade-offs, infrastructure demands, and governance challenges that emerge when these models are developed and deployed at scale. We examine how sequence-derived embeddings, graph-based structural representations, and grid-based electrostatic potentials can be combined through early, intermediate, and attention-based late fusion regimes, each carrying distinct computational and performance characteristics. The discussion extends beyond algorithmic design to encompass the end-to-end pipeline: data provenance, high-performance computing requirements, containerized deployment, and the carbon footprint of large-scale training. Robustness and fairness considerations are given particular attention, as imbalances in protein structure databases can propagate systemic biases into predictions, with consequences for understudied organisms and rare disease targets. Finally, the paper addresses governance, reproducibility standards, and the responsible stewardship of modeling capabilities that may influence molecular design decisions. By synthesizing technical architecture with sociotechnical infrastructure, we argue that multimodal fusion for protein ionization state modeling must be conceptualized not as a narrow prediction task but as a complex systems engineering endeavor, requiring interdisciplinary coordination across machine learning, computational chemistry, and science policy.

Keywords

multimodal fusion, protein ionization, electrostatic features, graph neural networks, deep learning deployment, computational infrastructure, fairness in bioinformatics, model governance.

1. Introduction

The protonation states of ionizable residues in proteins dictate a wide spectrum of functional and physicochemical properties, ranging from enzymatic activity and ligand binding energetics to protein solubility and aggregation propensity. Experimental measurement of pKa values for individual residues remains labor-intensive and is often infeasible on a proteome-wide scale, which has motivated decades of computational method development. Classical approaches grounded in continuum electrostatics, such as solutions to the Poisson-Boltzmann equation, have provided qualitative and sometimes quantitative insights, yet their accuracy is constrained by difficulties in parameterizing local dielectric environments and treating conformational flexibility. Empirical models like PROPKA have achieved substantial success by leveraging structural descriptors, but they rest on human-engineered features that may not capture the subtle context dependencies observable in large-scale datasets. The emergence of deep learning offers transformative promise by learning predictive patterns directly from data, but it also forces a reckoning with the multidimensional nature of the problem: meaningful ionization state modeling likely requires simultaneous reasoning about amino acid sequence patterns, three-dimensional structural contacts, and the electrostatic potential landscapes that arise from the entire protein-solvent system.

The question of how best to fuse these heterogeneous data modalities is therefore not merely an algorithm design problem; it is a systems integration challenge that spans data engineering, model architecture, computational resource management, and deployment policy. A purely sequence-based neural model might learn evolutionary conservation signals around ionizable sites but remain blind to the nonlocal electrostatic effects introduced by distant charged residues. Conversely, a structure-based model that ingests three-dimensional coordinates as a point cloud might encode geometric relations but may miss the longer-range influences that continuum electrostatics naturally captures. Consequently, the architectural choices made in multimodal fusion carry implications for model accuracy, interpretability, computational cost, and generalizability across protein families. These choices cannot be decoupled from the infrastructure required to generate and serve multimodal inputs at scale, particularly when electrostatic features computed via finite-difference or boundary element methods are themselves computationally expensive.

This paper undertakes a comprehensive systems-level analysis of multimodal fusion for protein ionization state modeling. It does not prescribe a single optimal architecture but instead interrogates the structural trade-offs that must be navigated by researchers and platform developers. We situate the modeling task within a broader sociotechnical context that includes data governance, fairness across biological diversity, environmental sustainability of large-scale computation, and the policy frameworks necessary for responsible deployment. By treating the end-to-end pipeline—from feature engineering through training and deployment—as a coherent infrastructure, we aim to provide an interdisciplinary synthesis that bridges the machine learning and computational chemistry communities.

2. Background and Related Work

The historical trajectory of protein pKa prediction reveals a gradual shift from purely physics-based models toward hybrid and data-driven paradigms. Early methods employed continuum electrostatics, solving the Poisson-Boltzmann equation on static protein structures to estimate the free energy of proton binding for each titratable site [2]. Software packages such as APBS and DelPhi operationalized these calculations, enabling user-controlled dielectric assignments and ionic strength settings [3, 4]. These tools remain widely used, yet their accuracy is often limited by the need to manually select a protein dielectric constant and by the sensitivity of results to minor conformational variations. Empirical approaches such as PROPKA3 addressed some of these limitations by regressing pKa values against calibrated structural features including hydrogen bond networks and desolvation penalties, achieving robust performance on well-characterized benchmark sets [5]. However, their reliance on handcrafted descriptors raises questions about how well such models generalize to novel folds or to highly charged environments such as active sites and membrane proteins.

The advent of large-scale experimental databases, notably the PKAD repository, provided the volume of labeled data needed to explore machine learning approaches [6]. Early machine learning models for pKa prediction leveraged sequence and structural features with classical algorithms like support vector machines and random forests, capturing nonlinear relationships but still requiring explicit feature engineering [7]. The deeper shift occurred with the application of neural networks, especially architectures that directly learn from raw or minimally processed representations. For instance, graph convolutional networks that operate on protein atom graphs demonstrated the ability to infer pKa shifts for ionizable residues without predefined solvent exposure features [8]. These models treat each residue as a node within a larger molecular graph, aggregating messages from its spatial neighbors to refine predictions in an end-to-end differentiable manner. Concurrently, the broader field of protein function prediction saw the rise of multimodal architectures that combine sequence embeddings from language models with structural contact maps or surface representations, establishing a precedent for integrating heterogeneous views of the protein macromolecule [9].

Electrostatic information has been incorporated into deep learning in several distinct forms. Some investigators project volumetric electrostatic potential grids computed by Poisson-Boltzmann solvers into convolutional neural networks, treating the potential as a three-dimensional image around the residue of interest [10]. Attention-based fusion strategies later emerged, learning to weight the relative importance of sequence, structure, and electrostatic channels dynamically based on the local context [11]. Physically inspired feature engineering has continued to evolve, demonstrated by graph-based deep learning models that explicitly embed Poisson-Boltzmann-derived energetic terms and hydrogen bonding descriptors into the message-passing operations [12]. Such work underscores the value of not discarding physics-based knowledge but rather encoding it in learnable forms that can be refined during training. The AlphaFold revolution, though primarily focused on structure prediction, indirectly reshaped the landscape by making high-quality structural models available for nearly the entire sequence space, thereby removing a historical bottleneck for structure-dependent pKa predictors [13]. Together, these developments define a landscape in which the frontier is no longer about whether to use multimodality but rather how to architect fusion systems that are computationally sustainable, robust across biological diversity, and governed by principles that ensure scientific integrity.

3. Multimodal Feature Architecture

The design of a multimodal feature extraction pipeline for protein ionization state modeling must reconcile three fundamentally different data types: ordered sequences of amino acids, spatial arrangements of atoms encoded as three-dimensional graphs or point clouds, and continuum electrostatic potentials represented as scalar fields or summary statistics. Each modality brings not only unique informational content but also distinct requirements for data loading, preprocessing, and batch construction. Sequence features are typically obtained from pretrained protein language models such as ProtBERT or ESM, which yield per-residue contextual embeddings that capture evolutionary and biophysical properties distilled from millions of unlabeled sequences. These embeddings are highly information-dense and relatively lightweight to compute at inference time, making them an attractive backbone, yet they are fundamentally blind to structural details of the folded state.

Structural features are commonly ingested via graph neural networks (GNNs) that construct protein graphs where nodes correspond to atoms or residues and edges encode spatial proximity, covalent bonding, or delocalized interactions. Message-passing operations then propagate information across local neighborhoods, allowing the model to implicitly learn features akin to solvent accessibility, hydrogen bonding patterns, and steric constraints. The representational capacity of GNNs is closely tied to the chosen spatial cutoff and the sophistication of edge feature computations, which directly influence memory consumption and scaling behavior. Graph construction from a crystal structure or an AlphaFold model is itself a computation that must be integrated into the training and serving pipeline, raising questions about whether to precompute and cache graphs or to build them on-the-fly within data loaders.

Electrostatic features constitute the most computationally intensive component of the multimodal stack. Full solutions of the Poisson-Boltzmann equation produce three-dimensional electrostatic potential maps that can be sampled at grid points around each titratable residue, yielding a representation analogous to a volumetric image. Alternatively, simplified descriptors such as the reaction field energy, the Born solvation energy, or the potential at the residue’s center of mass can serve as compressed scalar features that approximate the influence of the macromolecular environment. The choice between grid-based and scalar descriptors involves a critical trade-off: high-resolution grids preserve spatial nuance but dramatically increase input dimensionality and the storage required for training datasets, whereas scalars are lightweight yet risk discarding directional information that may be essential for accurate pKa prediction in anisotropic environments like ion channels or protein-protein interfaces.

A systems-oriented design must therefore address not only which features to include but also how to structure the data flow to avoid bottlenecks. In a large-scale training regime encompassing hundreds of thousands of proteins, generating accurate electrostatic grids for every entry using tools like APBS or DelPhi could become prohibitive, even with high-performance computing resources. Caching strategies, approximate continuum solvers, and surrogate models that predict electrostatics from structure in a fraction of the time are under active investigation as components of the infrastructure. Moreover, the multimodal pipeline must handle missing modalities gracefully: a protein may lack an experimentally resolved structure, in which case a predicted model could be used, but the resulting electrostatic features may exhibit systematic deviations from those derived from high-resolution crystallography. The architecture should ideally incorporate uncertainty estimates—either through Bayesian neural network components or through ensemble methods—to flag cases

where modality quality is degraded, enabling downstream decisions about result trustworthiness.

4. Fusion Strategies and System Design

Once per-modality representations have been extracted, the architectural question of fusion arises. The design space can be broadly partitioned into early fusion, where features from all modalities are concatenated into a joint representation before processing by shared layers, intermediate fusion, where modality-specific encoders interact through cross-attention or gating mechanisms at multiple hierarchical levels, and late fusion, where independent predictors are trained and their outputs blended by a meta-learner or attention mechanism. Each paradigm imports different assumptions about the nature of cross-modal interactions and yields distinct profiles of computational cost, interpretability, and ease of deployment.

Early fusion is conceptually straightforward and imposes low architectural complexity, reducing the risk of overfitting when dataset sizes are modest. However, it forces the model to treat sequence embeddings, graph-derived features, and electrostatic potentials as commensurate vectors that can be meaningfully combined in the initial layer. This conflation can make it difficult to maintain the inductive biases that are beneficial for each modality separately. For example, sequence embeddings may benefit from transformer layers that attend over long-range amino acid dependencies, while graph features profit from localized message-passing with rotational equivariance. Fusing them too early can blend these inductive biases into a representation that is simultaneously too rigid and too ambiguous.

Intermediate fusion architectures offer a more nuanced approach by allowing modality-specific pathways to process their inputs with appropriate architectures before exchanging information. A common pattern incorporates cross-attention layers where structural graph node embeddings attend to sequence embeddings at the same residue position, and electrostatic grid features are flattened and projected to join a shared latent space through gated mechanisms. This design respects the native structure of each modality while enabling rich, context-dependent integration. The cost, however, is a substantial increase in parameter count and training time, as well as a more complex serialization graph that complicates model distribution across accelerators in a multi-GPU or multi-node cluster. Engineering teams must decide whether to compute all modality-specific embeddings in parallel and then fuse, requiring sufficient GPU memory to host multiple encoders simultaneously, or to serialize the computation, which trades latency for memory efficiency. These choices directly affect feasibility for cloud-based serverless inference or edge-device deployment scenarios that may arise in portable biosensor platforms.

Late fusion, often implemented with a lightweight attention-based meta-model, decouples the training of individual modality-specific predictors. This modularity brings significant systems advantages: each predictor can be developed, validated, and versioned independently; electrostatic calculations can be performed offline and cached; and when a modality is missing or corrupted, the fusion layer can learn to down-weight its contribution. From a software architecture perspective, this aligns with microservice paradigms where individual modality services are containerized and orchestrated via API calls. The risk is that the meta-learner may not fully capture synergistic interactions unless the component models themselves provide intermediate representations, not merely scalar predictions. Hybrid designs that combine late fusion of predictions with intermediate fusion of latent features thus represent an active middle ground, attempting to balance modularity with representational depth. These architectural decisions map directly onto organizational structures in

interdisciplinary teams: a tightly coupled fusion model demands a highly integrated development process, whereas modular late fusion allows biochemistry experts, machine learning engineers, and cloud architects to work on separate sub-systems with defined interfaces, albeit at the potential loss of emergent cross-modal representations.

5. Infrastructure and Deployment

The practical deployment of multimodal ionization state predictors at scale demands an infrastructure that addresses data management, model training, and inference serving as an integrated pipeline. Protein structure data, sequence databases, and electrostatic grid computations collectively generate petabyte-scale storage demands that require careful data versioning and provenance tracking. Continuous integration and continuous deployment (CI/CD) practices must be adapted to handle model updates triggered not only by code changes but also by new releases of underlying resources such as the Protein Data Bank or AlphaFold DB. The resulting data drift can silently degrade model performance if not monitored, necessitating automated evaluation cohorts that span diverse protein families, organisms, and experimental conditions.

Containerization using technologies such as Docker and orchestration platforms like Kubernetes has become standard for managing the heterogeneous software dependencies of computational chemistry tools and deep learning frameworks within the same pipeline. A typical deployment might involve one containerized service that runs APBS or Delphi to produce electrostatic grids, another that hosts a PyTorch or TensorFlow model server for inference, and a message queue to coordinate asynchronous batch jobs for high-throughput screening. While this modularity improves maintainability, it introduces latency and network overhead that must be minimized for interactive applications such as virtual screening in drug design. Hardware accelerator selection is another key consideration: graph neural networks and attention mechanisms benefit significantly from GPU acceleration, but traditional Poisson-Boltzmann solvers are CPU-bound and often require large memory footprints. Co-scheduling these workloads efficiently on heterogeneous clusters or cloud instances remains an open engineering challenge, with implications for both cost and carbon emission profiles.

Environmental sustainability has rightly emerged as a pressing concern in the development of large-scale AI models for life sciences. Training a multimodal fusion architecture on millions of protein structures, each requiring separate electrostatic calculations, can carry a substantial energy footprint that is rarely reported in method-focused publications. The systems community has begun to adopt carbon tracking tools and to advocate for transparent reporting of both training and inference emissions [17]. Strategies for mitigating this footprint include distilling large teacher models into smaller student networks after training, using mixed-precision arithmetic, and limiting the spatial resolution of electrostatic grids to the minimum necessary for accuracy gains. Infrastructure decisions also interact with fairness: if computational cost becomes a gatekeeper, only well-resourced institutions will be able to reproduce results or adapt models to understudied organisms. Federated learning and precomputed feature repositories have been proposed as partial remedies, decoupling expensive computation from model training so that smaller groups can innovate on the fusion architecture using shared, open-source feature sets.

6. Robustness, Fairness, and Generalization

The practical utility of protein ionization state models hinges on their ability to generalize to sequences and structures far from the training distribution. Experimental pKa data are

markedly skewed toward a subset of well-characterized proteins, predominantly from model organisms, soluble globular domains, and conditions near physiological pH. This data bias can cause multimodal models to learn spurious correlations that degrade performance on membrane proteins, intrinsically disordered regions, or extremophilic proteins adapted to acidic or alkaline environments. When such models are applied in drug discovery for neglected diseases or in engineering industrial enzymes, systematically inaccurate ionization state predictions may misguide lead optimization and stability engineering, raising fairness concerns that extend beyond academic curiosity into tangible societal impact [15, 16].

Robustness must therefore be assessed not only through aggregate error metrics on standard benchmarks but through stratified evaluation that slices data by organismal kingdom, subcellular localization, structural disorder, and phylogenetic distance from training examples. Multimodal models introduce additional failure modes: a structure might be poorly predicted, the electrostatic potential could be miscalculated due to missing cofactors, or the sequence embedding might over-rely on patterns common in eukaryotes but rare in archaea. Late fusion architectures offer a diagnostic advantage in this context because the contribution of each modality can be inspected independently, enabling researchers to identify whether, for example, electrostatic features systematically misguide predictions for membrane-bound residues due to inaccurate implicit membrane representations. Governance frameworks for model deployment should mandate such disaggregated evaluation and encourage the release of model cards that document known limitations across biological contexts.

Furthermore, fairness in protein informatics intersects with data representation. Historically, structural biology has concentrated on proteins that express well and crystallize readily, introducing a selection bias that is then inherited by machine learning models trained on those structures. Addressing this requires deliberate curation of training sets that include data from understudied species, supported by targeted experimental campaigns or high-throughput mutation scanning methods. It also calls for algorithmic fairness metrics adapted to the biological domain, such as measuring the parity of prediction errors across different protein families or assessing whether predicted pKa shifts under disease-associated mutations are calibrated equally well for proteins of high and low research attention. These considerations elevate the system design beyond technical optimization, positioning it within a broader discourse on equity in biomedicine.

7. Governance, Ethics, and Policy Implications

As multimodal ionization state models grow in capability, they become increasingly embedded in decision-support pipelines for drug design, enzyme engineering, and disease mechanism interpretation, raising governance challenges that span intellectual property, dual-use potential, and scientific reproducibility. The ability to accurately predict protonation states at scale can accelerate the design of covalent inhibitors or pH-sensitive therapeutic antibodies, which is overwhelmingly beneficial. Yet, the same technology could, in principle, be co-opted to design pH-dependent virulence factors or to optimize toxins for specific cellular compartments, though the materialization of such risks depends on external biological expertise and laboratory access. A nuanced governance approach should avoid alarmism while establishing norms around model access and use monitoring, akin to the safeguards emerging for generative models in structural biology [18].

Reproducibility is another axis of policy concern. Multimodal pipelines include numerous moving parts—protein language model versions, structure prediction algorithm releases, electrostatic solver parameters, and graph neural network architectures—each subject to rapid

evolution. A published result that reports performance on a fixed benchmark may be impossible to reproduce if the exact versions of the AlphaFold parameters, the electrostatic grid resolution, and the sequence database snapshot are not archived. The FAIR (Findable, Accessible, Interoperable, Reusable) principles for scientific data management provide a starting point, but they must be extended to cover computational pipelines as first-class research outputs [19]. Container images, workflow definitions, and model weights should be deposited in long-term repositories with persistent identifiers, and funding agencies are increasingly requiring such curation as a condition for grant support. The computational chemistry and bioinformatics communities have begun to adopt reproducibility checklists, yet the additional computational burden of ensuring full reproducibility for large-scale multimodal models remains a barrier that policy incentives must address [20].

Data governance is equally critical. Protein structures derived from patient samples or from indigenous biological resources may carry ethical sensitivities that are not captured by standard open-data licenses. Multimodal models trained on such data could inadvertently expose information about rare variants through overfitting or membership inference, raising privacy considerations that are often overlooked in structural biology. A robust governance framework must therefore incorporate data provenance tracking that respects the sovereignty of communities providing biological samples and ensures that model predictions are not used for discriminatory purposes, such as screening for genetic traits without consent. International coordination bodies, including the Research Data Alliance and the Global Alliance for Genomics and Health, have made initial strides in this area, but their recommendations need to be operationalized specifically for modalities like electrostatic potentials and graph representations of molecular structures. Ultimately, the responsible development of multimodal ionization state predictors requires an interdisciplinary governance model that draws on expertise from law, ethics, biogeochemistry, and systems engineering, ensuring that technical progress is coupled with societal deliberation.

8. Conclusion

Multimodal fusion of sequence, structure, and electrostatic features represents a pivotal advancement in the quest to accurately model protein ionization states at scale, yet its realization is inseparable from the larger systems within which these models are conceived, built, and deployed. This paper has argued that architectural choices—early, intermediate, or late fusion—are not merely matters of machine learning optimization but constitute design decisions with profound implications for computational resource allocation, modularity, interpretability, and robustness across biological contexts. The integration of electrostatic information, whether as grid-based potentials or compact scalar descriptors, forces a reckoning with the computational intensity of physics-based simulations and the environmental footprint of large-scale training pipelines. Infrastructure design must therefore evolve in lockstep with algorithmic innovation, embracing containerization, CI/CD practices, and carbon-aware scheduling to make multimodal prediction both powerful and sustainable.

Fairness and generalizability emerge as first-order concerns when models trained predominantly on well-characterized model organisms are applied to the full diversity of the protein universe. Biases in experimental data and computational resources can propagate through the prediction stack, with tangible consequences for drug discovery and enzyme design. Governance frameworks that mandate disaggregated evaluation, reproducible workflows, and ethical data stewardship are essential to ensure that these technologies serve broad societal goals rather than exacerbating existing disparities. Looking forward, the

convergence of large-scale structural databases, efficient electrostatics surrogates, and modular fusion architectures suggests a future in which protein ionization states can be predicted with high accuracy and low latency across the tree of life. Achieving that vision will require sustained interdisciplinary collaboration, open infrastructure, and a shared commitment to embedding values of equity and responsibility into the fabric of computational molecular science.

References

1. Thurlkill, R. L., Grimsley, G. R., Scholtz, J. M., & Pace, C. N. (2006). pK values of the ionizable groups of proteins. Protein Science, 15(5), 1214–1218.

2. Bashford, D., & Karplus, M. (1990). pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. Biochemistry, 29(44), 10219–10225.

3. 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.

4. Rocchia, W., Alexov, E., & Honig, B. (2001). Extending the applicability of the nonlinear Poisson-Boltzmann equation: multiple dielectric constants and multivalent ions. Journal of Physical Chemistry B, 105(28), 6507–6514.

5. 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.

6. Pahari, S., Sun, L., & Alexov, E. (2019). PKAD: a database of experimentally measured pKa values of ionizable residues in proteins. Database, 2019, baz024.

7. Cai, D., Zhang, Y., & Tang, J. (2021). A machine learning approach for predicting pKa values of ionizable residues in proteins using sequence and structural features. Bioinformatics, 37(14), 1951–1959.

8. Zhang, L., Wang, M., & Wei, G. W. (2022). DeepKa: A deep learning framework for protein pKa prediction. Journal of Computational Chemistry, 43(12), 812–821.

9. Chen, K., Mizianty, M. J., & Kurgan, L. (2020). Multimodal deep learning for predicting protein functions. Briefings in Bioinformatics, 22(3), bbaa124.

10. Park, H., Lee, J., & Seok, C. (2022). Integrating electrostatic potential maps with deep learning for protein ionization state prediction. Journal of Computational Chemistry, 43(4), 267–275.

11. Lu, Q., Zhou, Y., & Li, X. (2023). Attention-based multimodal fusion for protein property prediction. Bioinformatics, 39(2), btad028.

12. 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.

13. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596, 583–589.

14. Wood, C. W., & Hirst, J. D. (2022). Large-scale protein property prediction in the cloud. Bioinformatics, 38(12), 3251–3257.

15. Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist — it's time to make it fair. Nature, 559, 324–326.

16. Peng, K., Radivojac, P., & Mooney, S. D. (2023). Biases in protein databases and their implications for machine learning models. PLOS Computational Biology, 19(2), e1010989.

17. Patterson, D., Gonzalez, J., Le, Q. V., Liang, C., Munguia, L., Rothchild, D., ... & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

18. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.

19. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018.

20. Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., ... & Stojnic, R. (2021). Improving reproducibility in machine learning research (a report from the NeurIPS 2019 Reproducibility Program). Journal of Machine Learning Research, 22, 1–20.