JCTC Journal of Chemical Theory and Computation

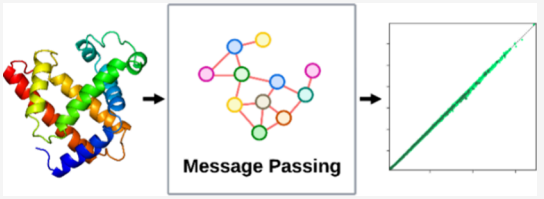
pubs.acs.org/JCTC Article

Accurate Predictions of Molecular Properties of Proteins via Graph Neural Networks and Transfer Learning

Spencer Wozniak, Giacomo Janson, and Michael Feig*

Cite This: J. Chem. Theory Comput. 2025, 21, 4830−4845

ABSTRACT: Machine learning has emerged as a promising approach for predicting molecular properties of proteins, as it addresses limitations of experimental and traditional computational methods. Here, we introduce GSnet, a graph neural network (GNN) trained to predict physicochemical and geometric properties including solvation-free energies, diffusion constants, and hydrodynamic radii, based on three-dimensional protein structures. By leveraging transfer learning, pretrained GSnet embeddings were adapted to predict solvent-accessible surface area (SASA) and residue-specific pK_a values, achieving high accuracy and generalizability. Notably, GSnet outperformed existing protein embeddings for SASA prediction and a locally charge-aware variant, aLCnet, approached the accuracy of simulation-based and empirical methods for pK_a prediction. Our GNN framework demonstrated robustness across diverse data sets, including intrinsically disordered peptides, and scalability for high-throughput applications. These results highlight the potential of GNN-based embeddings and transfer learning to advance protein structure analysis, providing a foundation for integrating predictive models into proteome-wide studies and structural biology pipelines.



INTRODUCTION

The three-dimensional (3D) structure of a molecule, typically represented by Cartesian coordinates, contains essential information for deriving its physicochemical and geometric properties.[1−3] Computer programs have long been used to calculate various molecular properties from structures,[4−6] especially properties that are unattainable or impractical to determine experimentally. This is particularly relevant for large biomolecules like proteins, where knowledge of properties like solvation free energy, hydrodynamic radius, solvent-accessible surface area, or pK_a can provide valuable insights into biological function and downstream tasks, like drug design.[7,8] Traditional methods, such as numerically solving the Poisson−Boltzmann equation to approximate solvation free energy[9] or running constant pH molecular dynamics (CpHMD) simulations to estimate pK_a values,[10] can prove to be computationally costly. Thus, these traditional methods may be inadequate for high-throughput predictions in the context of modern protein analysis and drug development.[8,11]

In recent years, the use of data-driven machine learning (ML) methods has emerged as an alternative solution for rapidly and accurately predicting molecular properties from structure. Such methods have been applied to small molecules[12] and proteins,[13] and they have been trained with experimental and/or simulated reference data.[14] While many of these models have demonstrated significant potential,[15−17] molecular ML methods, especially those designed for structural analysis of proteins, currently face many challenges. For one, developing a predictive model that is both accurate and generalizable requires large and diverse training data sets, but in practice, experimental data is often difficult to obtain, and computing reference data for a large set of molecules is expensive.[14] Moreover, molecular ML models are often designed to predict a narrow set of properties, so they are typically only useful for the specific tasks they were trained on, and predicting new properties generally requires training novel models with new training data sets.[18]

Transfer learning is a potential strategy to address these challenges. In transfer learning, a model first "learns" a latent representation (i.e., an "embedding") of input features that is optimized to predict target variables for which there is sufficient reference data.[19] Then, the learned embeddings of a "pretrained" model can be adapted to novel challenges for which training data is sparse,[20] like predicting pK_a values in small molecules.[21] By leveraging insights gained through pretraining, the model may be able to overcome constraints imposed by limited training data in the novel task.[22,23] "Learning" in this context is similar to latent learning in psychology, where knowledge is acquired without immediate application but becomes apparent through new challenges.[24]

In our work, we construct graph neural network (GNN) models[25] to produce latent representations of 3D protein structure via pretraining on supervised biomolecular property-

Received: December 10, 2024
Revised: March 14, 2025
Accepted: April 15, 2025
Published: April 24, 2025

https://doi.org/10.1021/acs.jctc.4c01682
J. Chem. Theory Comput. 2025, 21, 4830−4845

prediction tasks. GNNs have previously been shown to be well-suited for molecular data,[26−28] and they have been employed for producing molecular embeddings.[13,29] However, in previous work, pretraining was either accomplished via self-supervised training to predict relatively simple geometric features, like interatomic distances and bond angles,[13] or via contrastive learning to add higher resolution to coarse-grained protein structures.[30] Different from those approaches, we trained a global structure embedding network (GSnet) and an atomic local charge-aware embedding network (aLCnet) to produce embeddings that capture structural determinants of complex physicochemical features at different levels of resolution.

To obtain large data sets for pretraining, we leveraged ML methods for accurately predicting 3D protein structures from sequence.[31,32] We primarily used 3D protein models from the AlphaFold Protein Structure Database[33] and calculated molecular properties using standard physics-based approaches. Specifically, we trained GSnet to predict the radius of gyration (R_g), hydrodynamic radius (R_h), molecular volume (V), translational diffusion constant (D_t), rotational diffusion constant (D_r), and solvation free energy (ΔG_{sol}) of a protein from its 3D structure. The resultant network was able to predict these target molecular properties with high accuracy, and prediction accuracy remained high when applied to experimental structures from the protein data bank (PDB) and intrinsically disordered peptides (IDPs), even without including such proteins in the training set, demonstrating the broader transferability of our network.

We then applied this pretrained model to the prediction of molecular solvent-accessible surface area (SASA). Notably, the model was not originally trained to predict SASA, yet it predicted this property with high accuracy by leveraging the learned representations from pretraining. This outcome is notable when compared to other existing protein embedding models, such as GearNet,[13] another GNN that generates structural protein embeddings, and ESM-2,[32] a large language model (LLM) trained on extensive protein sequence databases that has been shown to excel in several structure prediction tasks, including protein structure prediction at atomic resolution. Neither of these alternative embeddings demonstrated similar success in SASA prediction, highlighting a unique advantage of GSnet.

Building on this, we next applied our model to predict pK_a values of amino acid residues within protein structures. Previous studies have utilized representation learning for pK_a predictions with small molecules[21] and proteins,[34] but GSnet and aLCnet embeddings are more general as, in principle, their learned representations could be leveraged for predicting other molecular properties. The GSnet and aLCnet models allowed for rapid predictions with similar or better accuracy than previously proposed physics-based and ML-based pK_a predictors.[34−36] Our best predictor approaches an accuracy of 0.9 pK_a units, and since the underlying GNN-based network is relatively lightweight, it is possible to rapidly predict ionization states of amino acids, even in very large complexes, or for large numbers of protein structures appropriate for proteome-scale annotation using experimental or modeled structures as input. We again compared our embeddings with GearNet[13] and ESM-2,[32] but we were unable to exceed the performance of the null model for pK_a prediction utilizing these other embeddings, demonstrating that GSnet and aLCnet embeddings capture richer information relevant to pK_a prediction than these other methods.

METHODS

Data Sets. GSnet was pretrained on the Swiss-Prot subset of the UniProt KnowledgeBase (UniProtKB/Swiss-Prot), utilizing protein structures as predicted by AlphaFold.[31,33] Of the 542,378 proteins in this data set, we only calculated reference values for a random subset of 153,513 of them due to considerable computational demands. We utilized HYDRO-PRO[37] to compute reference values for hydrodynamic radius (R_H), translational diffusion constant (D_t), rotational diffusion constant (D_r), and volume (V) using an atomistic representation. Radius of gyration (R_g) was calculated using the MDTraj library in Python[38] for the same subset of proteins.

We utilized the APBS software suite to compute solvation free energies (ΔG_{sol}) of proteins from PDB structure according to the linearized Poisson−Boltzmann equation.[5] PQR files, which include charges, were generated via PDB 2PQR using the CHARMM c36 force field.[39] The grid spacing was set to 0.15 Å, and dimensions were set such that the box was 10% wider than the protein in all three spatial dimensions. Because of the high computational expense of running APBS, ΔG_{sol} calculations were only obtained for a subset of 30,114 proteins out of the set of 153,513 for which we had values for hydrodynamic properties.

The entire data set was randomly split into training and validation sets at an approximate ratio of 9:1. The target values in both sets were normalized according to the mean and standard deviation of the data in the training set. Altogether, the training set consisted of 138,290 total structures ranging from 16 to 1,015 residues long, while the validation set consisted of 15,223 total structures ranging from 16 to 966 residues long.

Another data set consisting of 123 protein structures was used as a test set. This set was obtained via PISCES[40] and contained 123 dissimilar (<20% identity), small PDB structures, all of which had a resolution less than 1.5 Å.

Because our model was trained on AlphaFold-predicted structures, which inherently incorporate information from multiple sequence alignments, we also constructed an additional data set containing 23 orphan protein structures with no detectable sequence homologues. This data set was based on the Orphan25 data set proposed by Wang et al., which contains proteins that do not have any homologues in the Uni-Ref50_2018_03 data set.[41] Structures were obtained from the PDB, and reference values were calculated the same as with the other data sets. We omitted 2 structures from the data set, 7LOK_I and 7A5P_U, due to missing and/or nonstandard residues in the PDB structures.

To evaluate the generalizability of GSnet to alternative structure predictions, we also tested the model on subsets of approximately 100 training and 100 validation proteins, for which we generated ESMFold-predicted structures instead of AlphaFold structures. To create a representative sample for the training and validation sets, proteins were divided into three length categories: short (16−200 residues), medium (201−500 residues), and long (501+ residues). Within each length category, proteins were further stratified based on molecular volume quartiles to ensure a diverse range of structural properties. From each stratification bin, proteins were randomly selected in approximately equal proportions to guarantee that both short and long proteins, as well as proteins with different shapes, were included in the evaluation. Target values were computed for the ESMFold structures using the same methods as with the original AlphaFold structures. The GSnet model was

https://doi.org/10.1021/acs.jctc.4c01682
J. Chem. Theory Comput. 2025, 21, 4830−4845

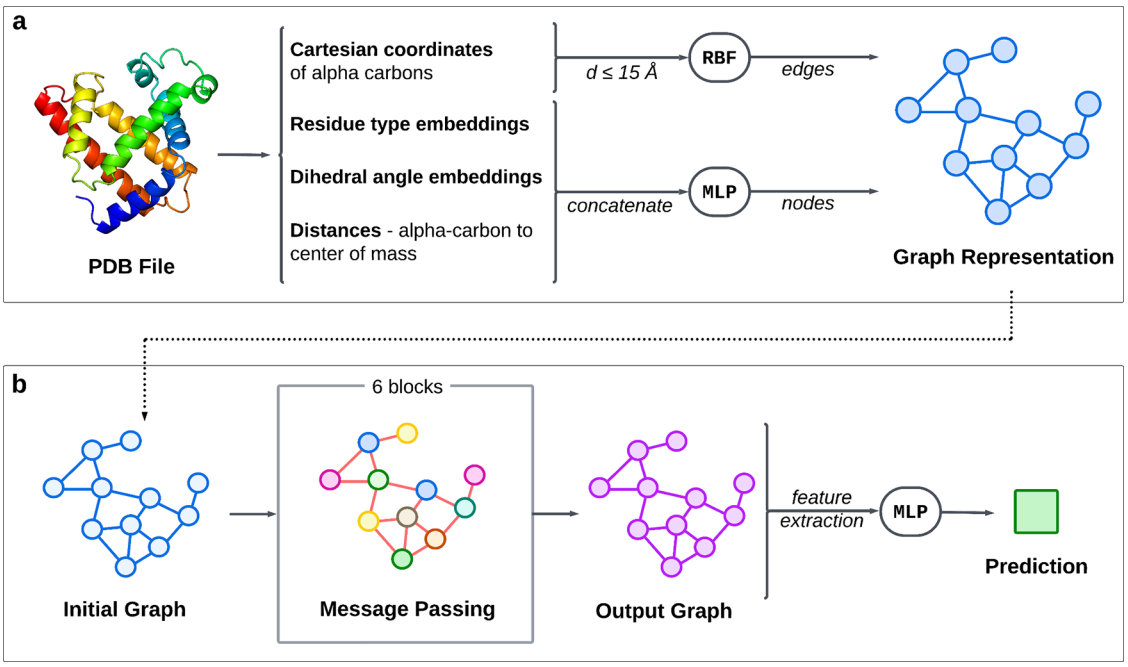


Figure 1. Overview of GSnet architecture. (a) Graph construction. Input node and edge features are extracted from the PDB structure to create a graph representation of the protein via multilayer perceptrons (MLPs) and a radial basis function (RBF) (see Figure S2 for more detail). **(b) Neural network architecture.** This initial graph is passed to a GNN consisting of six message-passing blocks (see Figure S4 for more detail), which ultimately generates an output graph with a high-dimensional embedding for each node. Features are extracted from the output graph and passed to a multilayer perceptron (MLP) to make final predictions.

then evaluated using the same input feature processing and normalization parameters as with the AlphaFold data set.

Another test set consisting of exclusively intrinsically disordered peptides (IDPs) was constructed, which contained ensembles of 100 structures for 45 distinct peptides (4,500 total structures). These ensembles were generated via COCOMO coarse-grained simulations,[42] followed by all-atom reconstruction via cg2all.[43] We also constructed a training set consisting of the ensemble of 100 structures for the shortest IDP, angiotensin, to fine-tune the model for predictions on IDPs.

For training on molecular solvent accessible surface area (SASA), we calculated reference values using the MDTraj library[38] with atomistic resolution for the same set of 153,513 protein structures described above, with identical splitting into training and validation sets.

For fine-tuning the GSnet to target residue-level SASA (rSASA) values, we calculated reference values with the MDTraj library[38] for a random subset of 259,049 residues from AlphaFold2 models for sequences in the UniProtKB/Swiss-Prot database,[31,33] with random splitting into training and validation sets at an approximate ratio of 9:1.

For making pK_a predictions, we utilized the PHMD549 data set, consisting of pK_a data obtained via constant pH molecular dynamics (CpHMD) simulations, and the EXP67S data set, consisting of 167 experimental pK_a data points, both of which were proposed by Cai et al.[36] We also constructed our own data sets consisting entirely of experimental pK_a values based on data obtained from PKAD-1[44] and PKAD-2,[45] as well as from the experimental references listed in Chen et al.,[35] Gokcan et al.,[34] and Wilson et al.[46] The sources of our data are outlined in Figure S1. For pK_a predictions, experimental PDB structures were obtained from the Protein Data Bank according to PDB codes provided with the data sets.

Initially, all pK_a data from PKAD-1 were assigned to a set. Nonidentical data points outside of PKAD-1 were then compared against all entries in this set. If a data point was structurally similar to (see below) any point in the set, it was added to this set. Otherwise, it was assigned to a separate set. This process produced two nonoverlapping data sets: a larger set of 1,932 total points and a smaller set of 237 total points.

The larger set (1,932 points) was split into training and validation sets at an approximate ratio of 9:1, yielding a training set of 1,738 entries and an initial validation set of 194 entries. Within the training set (MSU-pK_a-training), structurally similar data points were allowed, as this did not impact model evaluation. However, for the validation set, any similar entries were removed, leaving 52 unique data points in the final validation set (MSU-pK_a-validation).

The smaller set (237 points) was used as the basis for an independent test set. To ensure the test set contained only unique data points, we removed any structurally similar entries within this set, resulting in a final test set (MSU-pK_a-test) containing 143 unique data points. This process not only ensured uniqueness of each data point in the test set, but it also ensured that the test set remained fully independent of the training and validation sets, preventing data leakage in the construction of the MSU-pK_a-test, as seen in other studies and noted by Cai et al.[36] We note that our protocol for preventing data leakage may be more stringent than protocols in similar efforts described in recent papers,[47,48] where the same residues in proteins with highly similar sequences appear to be considered nonredundant,[47] or where the same residues in the same structures but with different PDB codes are present in published training and test sets.[48]

Similarity Classification Strategy. Structural similarity between proteins from the different training and test sets used in our study was measured via TM-scores. Each protein pair was

https://doi.org/10.1021/acs.jctc.4c01682
J. Chem. Theory Comput. 2025, 21, 4830−4845

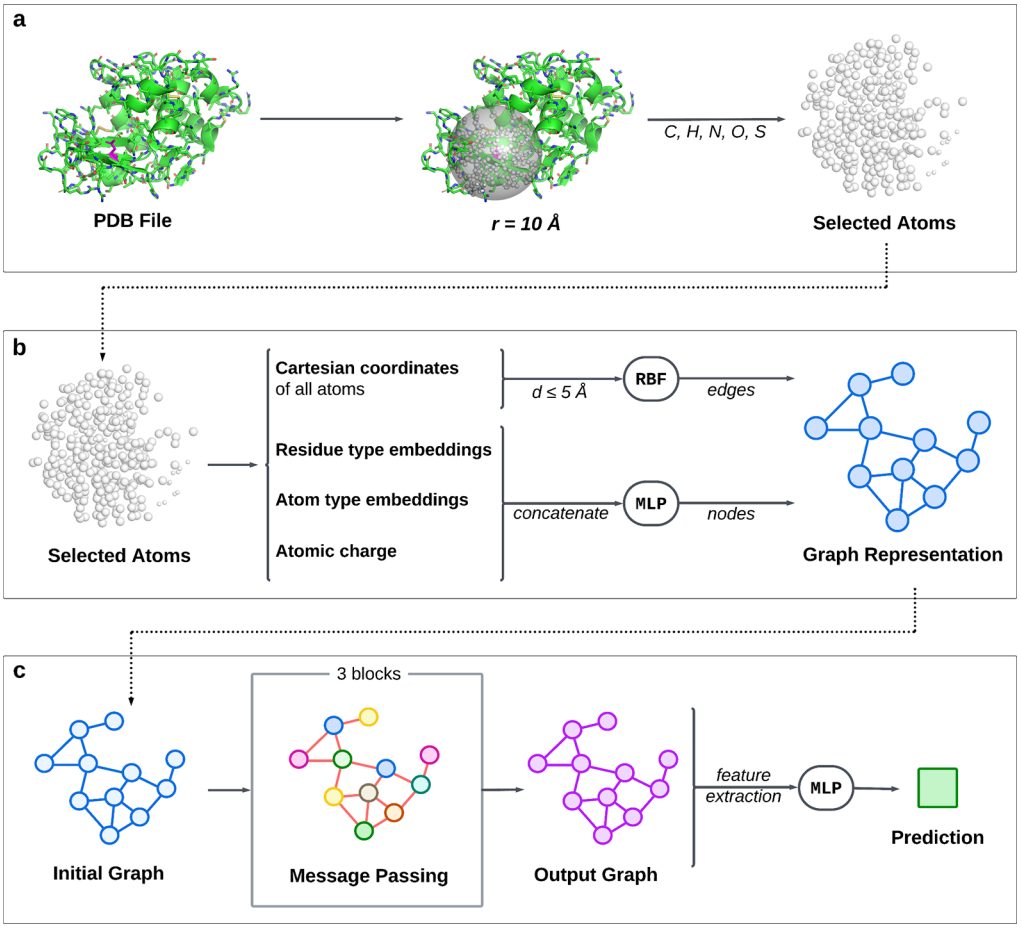


Figure 2. Overview of aLCnet architecture. (a) **Atom selection.** Carbon, hydrogen, nitrogen, oxygen, and sulfur atoms within ten angstroms of the α carbon of the residue of interest are selected from the original PDB structure. (b) **Graph construction.** Input node and edge features are extracted from the selected atoms to create a graph representation of the protein environment via a multilayer perceptron (MLP) and radial basis function (RBF), respectively (see Figure S3 for more detail). (c) **Neural network architecture.** Like GSnet, the initial graph is passed to a GNN, this one consisting of three message-passing layers (see Figure S4 for more detail), which ultimately generates an output graph with a high-dimensional embedding for each node. Features are extracted from the output graph and passed to a multilayer perceptron (MLP) to make a final prediction.

aligned with TM-align.[49] The average of the TM-scores were normalized by the lengths of the two proteins for which they were calculated. This average TM-score is reported and used throughout our analysis.

To determine sequence similarity within the pK_a data sets, 334 total sequences from PKAD-1, PKAD-2, and other sources were clustered using CD-HIT with a sequence identity threshold of 0.5, yielding 71 clusters. We then aligned the sequences within each cluster using ClustalOmega.[50]

A data point (i.e., an ionizable residue) was considered "similar" to another point if all of three conditions were met:

1. The parent sequence was found in the same cluster as the other data point.
2. The residue was in the same alignment column as the residue of the other data point.
3. The residue was of the same amino acid type as the residue of the other data point.

This strategy was used to remove structurally similar entries within the initial validation and test sets. This procedure also ensured that the training set S_{Train} and test set S_{Test} did not contain any data points corresponding to structurally equivalent residues in proteins with high sequence similarity. Specifically, for each data point $x \in S_{\text{Train}}$, we ensure there exists no corresponding data point $y \in S_{\text{Test}}$, such that the following conditions hold simultaneously:

$$\forall\, x \in S_{\text{Train}},\ \forall\, y \in S_{\text{Test}},\ \neg\,(P(x, y) \wedge A(x, y) \wedge T(x, y)) \tag{1}$$

where:

- $P(x, y)$ is true if proteins corresponding to data points x and y have a sequence identity greater than 0.5,
- $A(x, y)$ is true if the residues corresponding to data points x and y are in the same alignment column, and
- $T(x, y)$ is true if the residues corresponding to data points x and y are of the same amino acid type.

We also verified this condition was satisfied between the PHMD549 data set and the EXP67S data set, and between the PHMD549 data set and MSU-pK_a-test, to prevent data leakage during transfer learning. We found that there was 1 data point in the PHMD549 data set, 4O6U_A:134:HIS, that was similar to a data point in MSU-pK_a-test, 1B2 V_A:133:HIS, so it was not used in pretraining aLCnet.

Input Features & Graph Construction. An overview of the construction of GSnet is shown in Figure 1 and input features are listed in Table S1. For each protein in a data set, an initial graph $G = (V, E)$ was constructed with E(3) invariant

https://doi.org/10.1021/acs.jctc.4c01682
J. Chem. Theory Comput. 2025, 21, 4830−4845

features, with nodes V representing the residues within the protein and edges E between any residues with alpha carbons within 15 Å of each other. The node features $\mathbf{h}_i \in \mathbb{R}^d$ for each node $i \in V$ incorporated high-dimensional amino acid embeddings, information about dihedral angles, and the distance (in Å) between the constitutive Cα and the protein center of mass.

The amino acid embeddings were generated via the torch.nn.Embedding class in PyTorch, and they were learned during model training, with the aim of capturing relevant properties of the amino acids in the context of overall protein structure. The dihedral angle information incorporated the φ and ψ angles, as well as the first three χ angles (where applicable). This dihedral information incorporated both sine and cosine encodings, as well as a mask (0 if the angle does not exist; 1 if the angle exists), totaling 15 values per node (i.e., residue). We included dihedral angles as node features to provide the model with information about the spatial configuration of a protein's backbone and side chains, which should effectively provide the model with higher resolution. A detailed visualization of these dihedral angles can be found in Figure 1 of Chakrabarti and Pal.[51] The distance (in Å) between the constitutive Cα and center of mass of the protein was used as a node feature with the aim of providing the model information about each residue's location within the overall protein structure. Multilayer perceptrons (MLPs) were trained to learn high-dimensional tensor representations of nondiscrete input data, including the dihedral information and distances, and they were also employed for dimensionality reduction of input node features after concatenation (see Figure S2).

Edges E of the graph were generated between any residues that were within 15 Å of each other, with distances d_{ij} (in Å) calculated between the Cα atoms of residues i and j that represented an edge's corresponding nodes:

$$d_{ij} = \left\| \mathbf{p}_i - \mathbf{p}_j \right\| \tag{2}$$

where $\mathbf{p}_i$ and $\mathbf{p}_j$ are the Cartesian coordinates of the Cα atoms of residues i and j, respectively. We then applied Gaussian smearing to transform these scalar distances into higher-dimensional edge features $\boldsymbol{e}_{ij}$ according to

$$\boldsymbol{e}_{ij} = \left[\exp\left(-\frac{(d_{ij} - \mu_k)^2}{2\sigma^2} \right) \right]_{k=1}^{K} \tag{3}$$

where μ_k are the centers of linearly spaced Gaussian basis functions, σ is the spacing between consecutive μ_k values, and K is the total number of Gaussians used (i.e., the dimensionality of the edge features). We set $K = 300$, with μ_k linearly spaced from 0.0 to 15.0 Å in increments of approximately 0.05017 Å. Thus, the resulting edge features $\boldsymbol{e}_{ij}$ are 300-dimensional vectors. An overview of the extraction of GSnet node and edge embeddings from input features is shown in Figure S2.

An analogous graph construction was performed for aLCnet, shown in Figure 2 with input features listed in Table S2. When using aLCnet, an initial graph $G = (V, E)$ was also constructed with E(3) invariant features, but with nodes V for each carbon, hydrogen, nitrogen, oxygen, and sulfur atom within a 10 Å radius surrounding and including the alpha-carbon of the residue of interest, rather than for every residue in an entire protein. Here, node features $\mathbf{h}_i \in \mathbb{R}^d$ for each node $i \in V$ incorporated the same amino acid embeddings as GSnet, along with similar embeddings for atom type, as well as atomic partial charge as determined via PDB 2PQR.[5] Residue and atom embeddings were included with similar reasoning to GSnet residue embeddings. Atomic charge was incorporated to provide the model critical information about electrostatic interactions, which directly influence pK_a. Like with GSnet, MLPs were trained to learn high-dimensional tensor representations of nondiscrete input data (partial charges in this case), and for dimensionality reduction of input node features after concatenation (see Figure S3). Edges E were generated between any atoms that were within 5 Å of each other, with the only edge feature $\boldsymbol{e}_{ij}$ again derived according to eqs 2 and 3, but where $\mathbf{p}_i$ and $\mathbf{p}_j$ are the Cartesian coordinates of atom i and j, respectively, and with μ_k linearly spaced from 0.0 to 5.0 Å in increments of approximately 0.0167 Å.

Neural Network Architecture. With GSnet, the initial protein graph $G = (V, E)$ with 150-dimensional node features $\mathbf{h}_i \in \mathbb{R}^{150}$ for each residue $i \in V$ is passed through six custom transformer message passing layers based on those described by Shi et al.[52] Each message-passing layer applies attention to update the node features by attending over the neighboring nodes in G.

Before message passing, edge features $\boldsymbol{e}_{ij}$ are transformed via:

$$\boldsymbol{e}_{c,ij} = \boldsymbol{W}_{c,e,2}\sigma(\boldsymbol{W}_{c,e,1}\boldsymbol{e}_{ij} + \boldsymbol{b}_{c,e,1}) + \boldsymbol{b}_{c,e,2} \tag{4}$$

where $\boldsymbol{\sigma}$ is the shifted softplus activation function defined as $\boldsymbol{\sigma}(\mathbf{x}) = \ln(\mathbf{1} + e^{\mathbf{x}}) - \ln(2)\mathbf{1}$, and $\boldsymbol{W}$ and $\boldsymbol{b}$ are parameters that are learnable by the model. This operation was necessary to transform 300-dimensional edge features $\boldsymbol{e}_{ij}$ to 150-dimensional edge features $\boldsymbol{e}_{c,ij}$, in congruence with the dimensionality of the node features. Note that the subscript c denotes the attention head index, but GSnet only employs one attention head.

Next, the node features $\boldsymbol{H}^{(l)} = \{\mathbf{h}_1^{(l)}, \mathbf{h}_2^{(l)}, \cdots, \mathbf{h}_n^{(l)}\}$ in the graph at layer l are updated through the attention-based message-passing mechanism described by Shi et al.[52] These updated node features $\hat{\boldsymbol{H}}^{(l)} = \{\hat{\mathbf{h}}_1^{(l)}, \hat{\mathbf{h}}_2^{(l)}, \cdots, \hat{\mathbf{h}}_n^{(l)}\}$ were then passed through an additional residual connection to arrive at the updated node features $\boldsymbol{H}^{(l+1)} = \{\mathbf{h}_1^{(l+1)}, \mathbf{h}_2^{(l+1)}, \cdots, \mathbf{h}_n^{(l+1)}\}$ at the next layer $l + 1$ according to

$$\mathbf{h}_i^{(l+1)} = \mathbf{h}_i^{(l)} + \mathrm{ReLU}(\mathrm{LayerNorm}(\boldsymbol{W}_R\sigma(\hat{\mathbf{h}}_i^{(l)}) + \boldsymbol{b}_R)) \tag{5}$$

where LayerNorm is the layer normalization operation described by Ba et al.[53] Figure S4 shows a diagram outlining the entire message passing process.

We evaluated different message passing architectures, namely Schnet,[54] EGNN,[55] and TransformerConv,[52,56] as well as various numbers of attention heads, layers, and hidden channels, and the best performance we could achieve was done with these layers and parameters.

For global property predictions, the element-wise mean $\boldsymbol{\mu}$ across embeddings $\mathbf{h}_i$ for all nodes $i \in V$, following message-passing, was computed according to

$$\boldsymbol{\mu} = \frac{1}{|V|}\sum_{i \in V} \mathbf{h}_i \tag{6}$$

where $|V|$ is the total number of nodes in the graph. This global mean was then passed to an output MLP consisting of four linear layers with 1024 hidden channels each, and three shifted softplus (SSP) activation layers. This output MLP was trained to output the six target values, or for molecular SASA, the one target value.

https://doi.org/10.1021/acs.jctc.4c01682
J. Chem. Theory Comput. 2025, 21, 4830−4845

To make residue-level SASA predictions, an architecturally identical output MLP was used as with molecular SASA; however, the 150-dimensional node embedding specific to the residue for which a prediction was being made $\mathbf{h}_i$ was passed to the MLP, rather than the component-wise mean $\boldsymbol{\mu}$ (see Table S3). To make per-residue pK_a predictions, a total of seven 150-dimensional features were extracted and concatenated, resulting in a 1,050-dimesnsional feature (see eqs 6-8, Table S3). These features included $\boldsymbol{\mu}$ and $\mathbf{h}_i$, as well as five element-wise means across nodes representing residues within a spatial radius r (in Å) from the residue of interest $\boldsymbol{\mu}_r$:

$$\boldsymbol{\mu}_r = \frac{1}{|V_r|}\sum_{v\in V_r}\mathbf{h}_v \tag{7}$$

where $V_r \subseteq V$ denotes the subset of nodes corresponding to residues within r of residue of interest and $|V_r|$ is the number of nodes in this subset. Specifically, we used radii 6, 8, 10, 12, and 15 Å. The resulting features were concatenated according to

$$\text{concat}(\mathbf{h}_i, \boldsymbol{\mu}_6, \boldsymbol{\mu}_8, \boldsymbol{\mu}_{10}, \boldsymbol{\mu}_{12}, \boldsymbol{\mu}_{15}, \boldsymbol{\mu}) \tag{8}$$

The resulting tensor was then passed to an output MLP which consisted of six linear layers with 1024 hidden channels, six dropout layers with 20% dropout, and five SSP activation layers.

The architecture for aLCnet followed a similar pattern as GSnet, but with 75-dimensional node features $\mathbf{h}_i \in \mathbb{R}^{75}$ for each atom in the selection and only three of the previously described transformer-based message-passing layers (see Figure S4), each with three attention heads.

For pK_a predictions, three 75-dimensional features were extracted and concatenated, namely the element-wise mean across all atoms in the selection $\boldsymbol{\mu}$ (calculated via eq 6), the embedding representing the Cα atom in the residue of interest $\mathbf{h}_{C\alpha}$, and the element-wise mean across nodes representing atoms in the residue of interest $\boldsymbol{\mu}_{aa}$ (calculated via eq 9):

$$\boldsymbol{\mu}_{aa} = \frac{1}{|V_{aa}|}\sum_{v\in V_{aa}}\mathbf{h}_v \tag{9}$$

where $V_{aa} \subseteq V$ denotes the subset of nodes corresponding to atoms that compose the residue of interest and $|V_{aa}|$ is the number of nodes in this subset. These features were concatenated according to eq 10:

$$\text{concat}(\mathbf{h}_{C\alpha}, \boldsymbol{\mu}_{aa}, \boldsymbol{\mu}) \tag{10}$$

The output MLP here was identical to the one used for pK_a predictions with GSnet, with the only difference being a 225-dimensional, rather than 1,050-dimensional, input.

Training. To train GSnet on the geometric properties, as well as ΔG_{sol} (for which reference values were only available for 30,114 out of 153,513 proteins in the training set), a mask was applied to set the gradients to zero for instances where ΔG_{sol} reference values were not available, ensuring that model weights remained unchanged by missing values while still being updated based on available target values.[57]

We trained 20 models for each type of GNN layer, totaling 60 models, using the PyTorch[58] and PyTorch Geometric libraries.[59] We utilized the Adam optimizer[60] with an initial learning rate of 1×10^{-4}, adjusted to 1×10^{-5} after 50 epochs by a scheduler. Models were trained to minimize the mean squared error (MSE) loss across the six target values with a batch size of 64. Training was terminated when there was no improvement in validation performance for ten consecutive epochs (Figure S5). The model with the lowest validation MSE loss (of 20) across the six targets was selected for further evaluation on the test sets and for transfer learning applications.

To fine-tune GSnet for intrinsically disordered peptide (IDP) predictions, we first loaded the weights and biases of the GNN as obtained via pretraining on the original six target values. We then trained from here in the same fashion as the original training runs, using the ensemble of 100 angiotensin structures as the training set.

GSnet was fine-tuned similarly for molecular SASA predictions, but the parameters of the pretrained GNN were kept fixed such that only the output MLP was trained. This was done to test the broader transferability of the original weights and biases.

For residue-level SASA predictions, we trained models using two approaches. In the first, we applied the original GSnet (as trained on the six original properties) with fixed weights and only trained an output MLP. In the second, we trained an output MLP, as well as the fine-tuned GSnet itself by allowing optimization of the GNN weights based on the residue-level SASA data set. Training on this data set was performed similarly to before, but training was terminated after ten epochs. In each approach, five models were trained from the same original parameters, and the "best" model was selected based on lowest root mean squared error (RMSE) on the validation set.

For pK_a predictions based on GSnet, we used four training approaches: the initial GSnet with fixed GNN weights, the initial GSnet with GNN weights allowed to be optimized, the fine-tuned GSnet for predicting residue-level SASA values with fixed GNN weights, and the fine-tuned GSnet for predicting residue-level SASA values with GNN weights allowed to be optimized. In all cases, an output MLP was trained to map GNN embeddings to predictions. For each approach, we trained 20 total models to minimize MSE loss across the training pK_a data set, utilizing the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 64. A 20% dropout was necessary to prevent overfitting and to optimize validation performance. These models were trained for 100 epochs, and their performance was evaluated based on RMSE on the validation set. The ten best performing models based on validation RMSE were further evaluated on the test set to generate statistics. This procedure was conducted for both the data sets proposed by Cai et al.[36] and for our data sets. An overview of the entire training process for GSnet is shown in Figure S6.

With aLCnet, we also utilized transfer learning for pK_a prediction, but we did not pretrain with the same data as with GSnet. Instead, we first trained 20 randomly initialized aLCnet models on the simulated data contained in the PHMD549 data set. The top 10 models based on validation RMSE were selected and evaluated on the EXP67S test set. We then selected the pretrained model with the lowest validation RMSE for fine-tuning using MSU-pK_a-training and MSU-pK_a-validation. Again, 20 training runs were performed and the top ten models were evaluated on MSU-pK_a-test. An overview of the entire training process for aLCnet is shown in Figure S7.

Alternative Embeddings. When training our GNN models on the original six values, molecular SASA, and pK_a, we compared their performance to the performance of ESM-2,[32] a large language model using sequences as inputs, and GearNet,[13] another structure-based GNN. The 650 million parameter ESM model that we used generated 1,280-dimensional embeddings for each residue, while the GearNet model generated 3,072-dimensional embeddings for each residue. For global property

https://doi.org/10.1021/acs.jctc.4c01682
J. Chem. Theory Comput. 2025, 21, 4830−4845

Table 1. Validation Set Performance for Prediction of Global Molecular Properties[a]

	Model	ΔG_{sol}[kJ/mol]	R_g[Å]	R_h[Å]	D_t[nm^2/μs]	D_r[μs^{-1}]	V [nm^3]
MAPE (%)	*Transformer (GSnet)*	**4.21 ± 0.04**	**0.97 ± 0.01**	**0.67 ± 0.00**	**0.65 ± 0.00**	**2.79 ± 0.03**	**1.10 ± 0.01**
		(3.89)	**(0.82)**	**(0.65)**	**(0.65)**	**(2.47)**	**(1.12)**
	EGNN	**4.15 ± 0.03**	1.09 ± 0.01	0.69 ± 0.01	0.68 ± 0.00	2.93 ± 0.04	1.13 ± 0.01
		(3.95)	(1.22)	(0.73)	(0.71)	(2.71)	(1.20)
	SchNet	5.71 ± 0.03	1.25 ± 0.02	0.81 ± 0.00	0.78 ± 0.00	3.41 ± 0.03	**1.11 ± 0.01**
		(5.84)	(1.15)	(0.80)	(0.80)	(3.60)	**(1.12)**
	GearNet	6.75 ± 0.03	3.05 ± 0.01	1.97 ± 0.00	1.97 ± 0.00	6.64 ± 0.02	4.07 ± 0.01
		(6.67)	(3.00)	(1.96)	(1.95)	(6.66)	(4.07)
	ESM-2[b]	9.04	5.02	2.91	2.91	10.2	5.43
		(9.04)	(5.02)	(2.91)	(2.91)	(10.2)	(5.43)
RMSE	*Transformer (GSnet)*	565.5 ± 2.4	**0.38 ± 0.00**	**0.31 ± 0.00**	**0.71 ± 0.00**	**0.38 ± 0.00**	**0.66 ± 0.01**
		(532.0)	**(0.35)**	**(0.30)**	**(0.71)**	**(0.37)**	(0.65)
	EGNN	**555.6 ± 2.8**	0.42 ± 0.00	0.32 ± 0.00	0.73 ± 0.00	0.40 ± 0.00	**0.66 ± 0.01**
		(538.0)	(0.42)	(0.31)	(0.77)	(0.39)	**(0.64)**
	SchNet	844.7 ± 2.1	0.47 ± 0.00	0.37 ± 0.00	0.84 ± 0.00	0.48 ± 0.00	**0.67 ± 0.01**
		(837.0)	(0.46)	(0.36)	(0.86)	(0.49)	(0.71)
	GearNet	1031.1 ± 3.0	1.65 ± 0.00	1.04 ± 0.00	2.20 ± 0.00	0.89 ± 0.00	3.51 ± 0.01
		(1032.4)	(1.62)	(1.03)	(2.18)	(0.88)	(3.58)
	ESM-2[b]	1298.0	2.24	1.38	3.29	1.56	3.77
		(1298.0)	(2.24)	(1.38)	(3.29)	(1.56)	(3.77)

[a]Mean absolute percent errors (MAPE) and root mean square errors (RMSE) are reported as the mean and standard errors from 20 independent training runs (except for ESM-2), with the value for the best overall model (chosen by lowest MSE across all target values) given in parentheses. Bold values indicate the best performance across all models for a given target value. [b]Only one model was trained with ESM-2 due to computational cost.

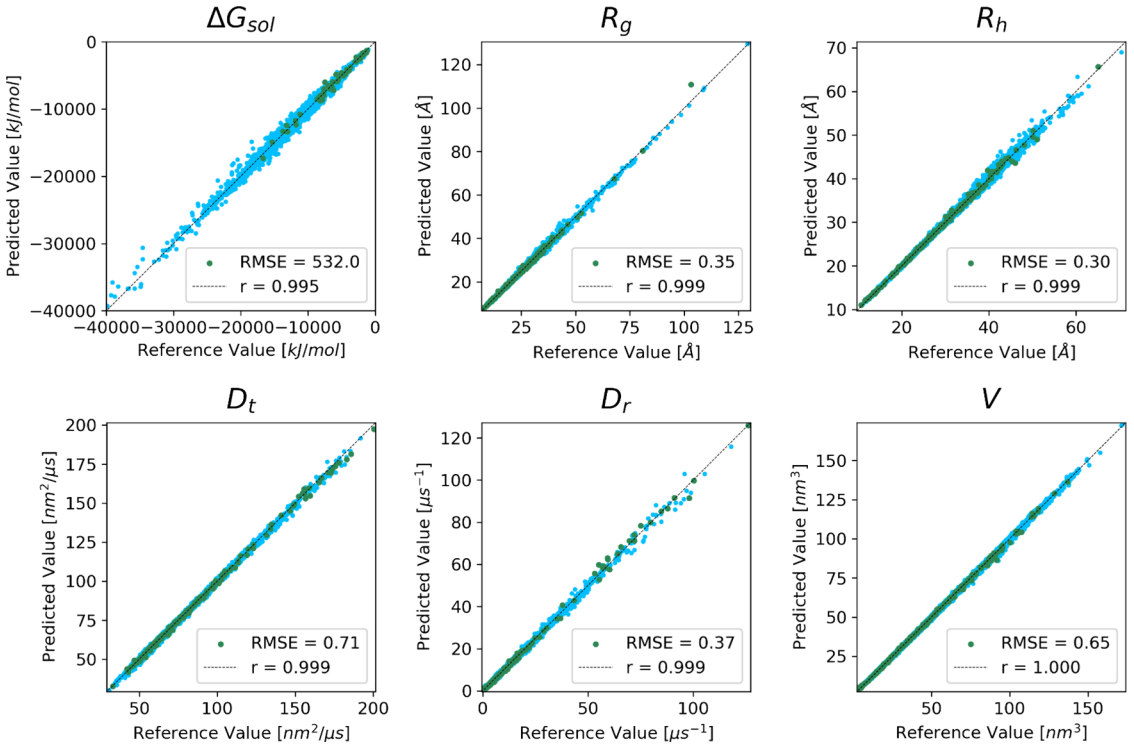


Figure 3. GSnet predictions of global molecular properties vs calculated reference values. Predictions and calculated reference values were made using structures as predicted by AlphaFold. Results are shown for the predicted global molecular properties in the validation set: ΔG_{sol}, R_g, R_h, D_t, D_r, and V. Darker green points represent structures that are not homologous to any structures in the training set. Pearson correlation coefficients and RMSEs encompass all data shown in the plots. RMSE values are given in units as displayed on the axes of each subplot.

predictions with these models, the mean of embeddings was calculated as in eq 6 and used as an input feature. For pK_a predictions, model embeddings were concatenated in the same fashion as GSnet embeddings (see eqs 6-8, Table S3) to obtain input features. These features were passed to identical output MLPs as were used in the different applications of our model, with the only difference being the number of input channels, based on the embedding dimension of the given model.

Code is available on GitHub at https://github.com/feiglab/ProteinStructureEmbedding

https://doi.org/10.1021/acs.jctc.4c01682
J. Chem. Theory Comput. 2025, 21, 4830−4845

RESULTS AND DISCUSSION

Global Structural Embeddings. To create global structure embeddings, a GNN and output MLP (Figure 1) were trained with the goal of simultaneously predicting six molecular properties of proteins, namely free energy of solvation (ΔG_{sol}), radius of gyration (R_g), hydrodynamic radius (R_h), translational diffusion coefficient (D_t), rotational diffusion coefficient (D_r), and molecular volume (V). These properties were selected based on their significance in characterizing protein structure and function: R_g is a geometric property that captures the distribution of atoms within the molecular cavity; hydrodynamic properties like R_h, D_t, and D_r capture the external shape of the protein and its interactions with the surrounding solvent; and ΔG_{sol} captures information about the charge distribution, allowing the inference of details about electrostatic interactions. The computational calculation of hydrodynamic properties and ΔG_{sol} via traditional software is very costly, taking up to days for a single protein, highlighting the need for an efficient and accurate ML model.

Various GNN implementations were tested, namely SchNet,[54] an EGNN,[55] and a transformer,[52,56] to determine which architecture provides the highest predictive accuracy for these properties. We found similarly high prediction accuracy across all three of these GNN architectures when applied to R_g, R_h, D_t, D_r, and V (Table 1), and validation set performance was only moderately worse than training set performance (Table S4). Among our GNN models, the transformer architecture performed slightly better, achieving the lowest mean absolute percent errors (MAPEs) and root mean squared errors (RMSEs) across the target values (Table 1). Only for ΔG_{sol}, the EGNN model was slightly better (Table 1). Therefore, we adopted the transformer model (referred to as GSnet) for transfer learning applications.

We also tested alternative pretrained ML models that generate protein representations, namely ESM-2[32] with sequences as input, and GearNet[13] with 3D structures as input. An architecturally identical output MLP (except for number of input dimensions, as dictated by the embedding size of the tested model) as used with our GNN models was trained to predict the six properties based on these alternative input embeddings. All our GNN models significantly outperformed networks based on GearNet and ESM-2 embeddings (Table 1).

The performance of GSNet for predicting global molecular properties is illustrated further in Figure 3. Strong correlation is found for the systems in the validation set across the entire range of values, only slightly worse than the correlation obtained for the training set (Figure S8). To confirm that the validation accuracy of the model did not result from data leakage due to homologous structures in the training set, we performed BLAST alignments[61] for each validation set protein against all proteins in the training set and established a threshold for homology at an E-value[62] of 10^{-5}. We found that the performance of the models on the validation set proteins that did not exhibit homology to any proteins in the training set was roughly the same as for the proteins that did, indicating that the accuracy was independent of data leakage (Figure 3).

Across all geometric properties, GSnet exhibits an error ranging from 0.65 to 2.47%, which is lower than HYDROPRO's error of approximately 4% relative to experimental values.[6] This suggests that, given enough accurate experimental training data, GSnet could surpass HYDROPRO's accuracy in predicting hydrodynamic observables. However, predictions of ΔG_{sol}, with an error of 3.89%, are significantly less accurate than continuum electrostatics calculations, where the typical error may be 0.25% or lower.[63,64] The relatively poor performance of GSnet for predicting ΔG_{sol} may result from limited training data relative to the other features, or because it is inherently more difficult to predict solvation free energies from structural embeddings alone.

Since GSnet was trained using AlphaFold (AF) models, we further evaluated GSnet on a test data set consisting of 123 structures of smaller proteins from the Protein Data Bank. Performance on this data set was generally similar to the performance of the model on the larger data set for R_g, D_t, and D_r (Figure S9). However, the model did have a slightly higher RMSE for ΔG_{sol}, R_h, and V on this PDB test set when compared with the AlphaFold validation set. In Figure S9, we stratified data points by greatest TM-score to any training set protein and observed no obvious discrepancies in performance between those with a maximum TM-score greater or less than 0.6, suggesting that fold-level similarity does not strongly influence model accuracy.

Notably, GSnet seemed to overestimate on ΔG_{sol} predictions relative to APBS on the PDB test set. Further evaluation of GSnet on a larger set of 610 PDB structures[63] showed the same tendency (Figure S10) across all PDB structures. To test whether this shift was due to issues with PDB structures, a brief energy minimization was conducted via CHARMM[65] and reference values were calculated again. On the minimized structures, GSnet underestimated solvation free energies relative to APBS (Figure S10). Subsequently, AF structure predictions for the same subset of proteins were obtained and reference values were calculated for these structures. Interestingly, the model predictions on these structures restored the initial accuracy reported for the validation set without systematic deviations in either direction. This suggests that GSnet learned specific features of AF models when making ΔG_{sol} predictions.

It has been shown that SASA of ionizable residues is greater with structures predicted by AF relative to experimental structures.[36] Thus, solvation free energy is likely to be lower (i.e., more negative) for AF structures relative to experimental structures, as increased solvent exposure would make solvation more thermodynamically favorable. It is possible that our model, which was trained on AF structures with ionizable residues that are too exposed, became biased during training, which led to discrepancies when applied to experimental structures. While this may not be a significant issue for the purpose of generating a reusable structural embedding in this work, future efforts may consider how to make predictions of solvation free energies with model-independent accuracies.

We also evaluated GSnet on a test set of 23 orphan proteins to assess its generalizability beyond proteins with known homologues. When evolutionary information is available, AlphaFold leverages multiple sequence alignments (MSAs) and templates for structure prediction, and previous studies have shown that AlphaFold models tend to have lower confidence and increased local structural deviations for orphan proteins.[41,66] Despite this, when compared to the small PDB test set, GSnet exhibited no significant reduction in prediction accuracy on this orphan protein set (Figure S11). As observed with the PDB test set, GSnet maintained similar performance for R_g, D_t, and D_r, while showing slightly worse performance on ΔG_{sol}, R_h, and V relative to the AlphaFold validation set (Figure S11). Our results suggest that GSnet does not rely heavily on evolutionary information embedded within AlphaFold training

https://doi.org/10.1021/acs.jctc.4c01682
J. Chem. Theory Comput. 2025, 21, 4830−4845

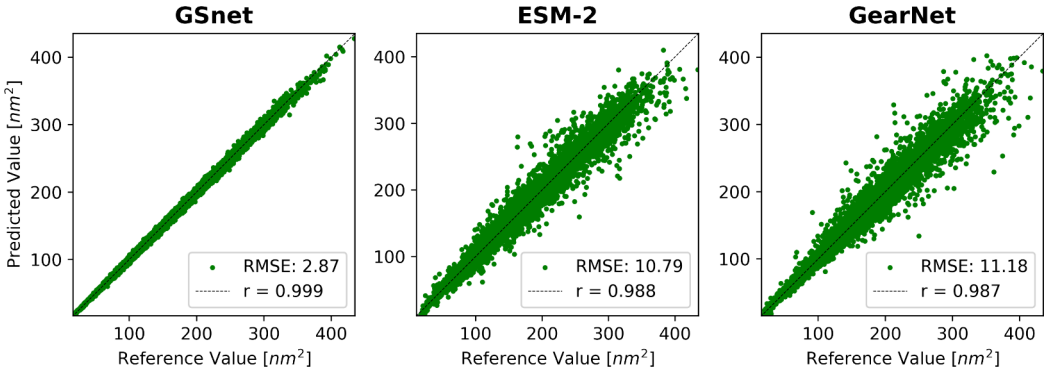


Figure 4. Validation set performance of embedding-based SASA predictors. Plots show predicted SASA values versus calculated reference values. Predictions and calculated reference values were made using structures as predicted by AlphaFold. Correlation coefficients from linear fits are shown in the upper left corner of each plot. RMSE values are given in nm^2.

structures. Its predictive accuracy remains consistent even for proteins without detectable sequence similarities to known proteins, indicating that GSnet effectively captures structural determinants of molecular properties in a way that generalizes beyond homologous protein families.

We also tested GSnet on subsets of 100 proteins from the original training and validation sets, using structures as predicted by ESMFold, to estimate the generalizability of GSnet to structures beyond AlphaFold and experiment (Figures S12 and S13). Overall, we find that GSnet maintains reasonable performance when using ESMFold structures, with predictions for most properties closely matching those obtained using AlphaFold structures. Similar results may be expected because both AlphaFold and ESMFold share the same structure module architecture for generating the final output. However, we observe somewhat higher RMSEs for ΔG_{sol} when using ESMFold structures (746 and 701 kJ/mol for the training and validation subsets, respectively) compared to the original AlphaFold-based results (311 and 532 kJ/mol). This suggests that ΔG_{sol} predictions may be more sensitive to inaccuracies in ESMFold-predicted structures than other molecular properties. Predictions for R_h also show a slight decline in accuracy. Interestingly, we observe that RMSEs for all properties tend to be lower in the validation subset than the training subset when using ESMFold structures. These findings suggest that GSnet is relatively robust to structural variations introduced by ESMFold.

Finally, another data set was considered to explore the transferability of GSnet toward intrinsically disordered peptides (IDPs). IDPs were generally not present in the training set, although some of the training set proteins certainly contained short, disordered elements such as loops as part of folded structures. The IDP data set consisted of conformational ensembles generated via coarse-grained simulations using COCOMO,[42] followed by all-atom reconstruction.[43] All reference values were obtained in the same manner as for the other data sets. GSnet predicted the target properties again with high correlation for this data set; however, performance was generally worse compared to folded proteins (Figure S14). The model exhibited a propensity to underestimate R_g, R_h, and V, while overestimating D_t and D_r (Figure S14). These shifts are likely due to the decreased compactness of IDPs as compared to folded proteins. Because radius of gyration, hydrodynamic radius, and volume inherently measure a protein's size and spatial occupancy, GSnet trained on folded proteins may be biased toward predicting lower values, similar to those observed in more compact, folded proteins. Accordingly, considering that diffusion coefficients are inversely related to protein size,[67] the model may anticipate greater values. Additionally, the model did not perform well on very small IDPs, notably angiotensin (8 residues), YESG2 (8 residues), and YESG6 (12 residues), all of which were shorter than any of the proteins found in the original training set. The performance of GSnet on the IDP set shows limitations of GSnet when transferred to structures unlike those in the training set.

To test whether additional training could improve the model, GSnet was fine-tuned with an additional ensemble of 100 angiotensin structures. In the fine-tuned model, most of the systematic shifts were removed and the large errors for very short peptides were eliminated (Figure S15). This indicates that modest data augmentation may be enough to adapt GSnet to structures outside the domain of the original training set.

Predicting of Global Properties via Transfer Learning from GSnet. To explore the transferability of GSnet embeddings to other global properties, we focused on the prediction of solvent-accessible surface areas (SASA). The parameters of the pretrained GNN were loaded and kept fixed during this process, while a new, trainable output MLP was introduced to map the element-wise mean of the GNN node embeddings for a given protein structure to its SASA value. After training the new MLP to predict SASA based on the pretrained GNN embeddings, on the same structures as in the original training set, the model showed excellent performance on the validation data (Figure 4). For comparison, we also considered the use of embeddings from ESM-2[32] (sequence-based) and GearNet[13] (structure-based). The training and validation performance were significantly worse when using either of those embeddings (Figure 4, Figure S16), indicating that GSnet may be a better platform for transfer leaning applications for the prediction of protein structure properties.

Interestingly, the validation set performance of GSnet-based SASA predictions is very similar to the training set performance (Figure S16). This is not the case for GearNet-based SASA predictions where validation set performance is significantly decreased over the training set performance, indicating better transferability with GSnet embeddings. To further explore the transferability of GSnet embeddings, we evaluated the model on three subsets of the original training set with proteins shorter than 50, 100, or 200 residues, respectively. The motivation for considering these subsets was to test whether transferability to larger proteins could be achieved without including such proteins during training, since the calculation of reference values needed for training becomes increasingly costly as the

https://doi.org/10.1021/acs.jctc.4c01682
J. Chem. Theory Comput. 2025, 21, 4830−4845

Table 2. Prediction of Experimental pK_a Values from the EXP67S Test Set[a]

Method	RMSE	R	m
Null Model[b]	1.44	-	0
PROPKA[b]	1.12	0.63	0.45
CpHMD[b]	1.01	0.73	0.65
PKAI+[b]	1.30	0.45	0.15
DeepKa[b]	0.97	0.74	0.50
GSnet[c]	1.39 ± 0.01 (1.37)	0.40 ± 0.01 (0.42)	0.14 ± 0.00 (0.15)
GSnet (rSASA)[d]	1.31 ± 0.01 (1.28)	0.52 ± 0.00 (0.53)	0.19 ± 0.00 (0.21)
GSnet (opt)[e]	1.33 ± 0.00 (1.31)	0.49 ± 0.01 (0.52)	0.19 ± 0.01 (0.23)
GSnet (rSASA-opt)[f]	1.29 ± 0.00 (1.26)	0.52 ± 0.00 (0.54)	0.24 ± 0.01 (0.28)
aLCnet (CHARMM)[g]	1.10 ± 0.01 (1.08)	0.63 ± 0.01 (0.67)	0.38 ± 0.01 (0.48)
aLCnet (AMBER)[g]	1.08 ± 0.01 (1.05)	0.62 ± 0.01 (0.66)	0.38 ± 0.00 (0.41)

[a]Performance with our GNN-based models is given as mean ± SEM with the best test RMSE in parentheses. Performance data for the null model, PROPKA, CpHMD, and PKAI+ was taken from from Cai *et al.*.[36] Performance data for DeepKa was obtained via predictions made with the DeepKa web server.[68] All predictions were made using experimental PDB structures. [b]values taken from Cai et al.[36] [c]using pretrained GSnet model with fixed GNN weights. [d]using pretrained GSnet model fine-tuned for residue-level SASA with fixed GNN weights. [e]using pretrained GSnet model with optimized GNN weights. [f]using pretrained GSnet model fine-tuned for residue-level SASA with optimized GNN weights. [g]trained from initialized weights (i.e., not pretrained).

protein size increases. In addition to strictly limiting protein size in the training set, we also considered a data augmentation strategy where a few larger proteins were added to a training set that was otherwise limited in size. Training on size-limited training sets, we found that the accuracy of predictions for proteins larger than the training set limit quickly deteriorated (Figure S17), but the performance was better with augmented training data (Figure S18). Specifically, training on proteins with up to 200 residues augmented by a few larger proteins resulted in fairly accurate validation set predictions (r = 0.997, RMSE = 6 nm^2) across the entire range of proteins up to 1000 residues (Figure S18). This suggests that transfer leaning based on GSnet does require that the training set domain covers the target application domain, but training of accurate models may be possible with sparser data than what was used in the original training of GSnet.

Predicting of Local Properties via Transfer Learning from GSnet. We also tested whether GSnet embeddings were useful in making residue-level predictions. We first applied the embeddings of GSnet to predict residue-level SASA values by training an output MLP with fixed, pretrained GNN parameters. We found relatively poor correlation (r = 0.751) and RMSE (0.36 nm^2), and the relative errors were significantly larger than for the predictions of the global structures (Figure S19A). We then tested whether fine-tuning the GNN itself (i.e., allowing the parameters of the pretrained GNN to be optimized along with the output MLP) can improve accuracy. We found significant improvements in correlation (r = 0.956) and RMSE (0.16 nm^2) relative to fixed GNN parameters, but the relative errors remained larger than for the prediction of global properties (Figure S19B), indicating that the GSnet embedding may not have enough capacity to predict residue-level properties with the same accuracy as global structural properties.

We then considered the prediction of residue-specific pK_a values. We first focused on using simulated pK_a data from the PHMD549 data set for training,[36] and then we evaluated the model on the EXP67S test set consisting of experimental pK_a values as in previous work.[36] We used simulated pK_a for training here to have enough training data to evaluate different models and training protocols, whereas testing the resulting models against experimental pK_a values allowed us to compare with other approaches from other groups. Here, we applied four training approaches: the initial GSnet with either fixed or optimized GNN weights, and the fine-tuned GSnet for residue-level SASA predictions, again with either fixed or optimized GNN weights.

As shown in Table 2, fine-tuning the model for residue-level SASA predictions before pK_a training resulted in more accurate pK_a predictions, with average RMSE decreasing from 1.39 to 1.31 when not optimizing GNN weights, and decreasing from 1.33 to 1.29 when GNN weights were allowed to optimize. This improvement suggests that fine-tuning on residue-level SASA predictions helped the model better capture local structural properties relevant to residue-specific properties, such as pK_a. Moreover, allowing the GNN to optimize during pK_a training led to marginal, but significant, improvements in accuracy, with average RMSE dropping from 1.39 to 1.33 with the original GSnet weights, and dropping from 1.31 to 1.29 with the fine-tuned GSnet weights on residue-level SASA. This improvement indicates that allowing GNN weights to optimize enables the model to refine its representations further to account for structural features uniquely relevant to pK_a prediction.

While the average performance of the doubly fine-tuned GSnet variant (RMSE = 1.29) exceeded the performance of the null model (RMSE = 1.44) and PKAI+ (RMSE 1.30), it fell short when compared to other methods like PROPKA (RMSE = 1.12), CpHMD simulations (RMSE = 1.01), and DeepKa (RMSE = 0.97). This analysis demonstrates that transfer learning targeting more complex residue-level properties based on structural embeddings is possible, but the performance achieved so far lags behind other methods that were specifically developed for pK_a predictions.

Local Charge-Aware Embedding for pK_a Predictions. Although GSnet was trained to reproduce geometric, hydrodynamic, and electrostatic solvation free energies, it does not explicitly consider information about local charge distributions which may be essential for making accurate pK_a shift predictions. To test whether a charge-aware local structural embedding can improve pK_a predictions, we developed aLCnet, an atomic variant of our GNN model that is more lightweight than GSnet and tailored specifically to pK_a predictions by including partial atomic charges as an input feature. We tested various edge cutoffs (radii around each node for drawing edges) and maximum numbers of edges per node and found that the best

https://doi.org/10.1021/acs.jctc.4c01682
J. Chem. Theory Comput. 2025, 21, 4830−4845

Table 3. Performance Metrics for Different Models Evaluated on MSU-pKa-test[a]

Method	RMSE	R	m
Null Model	1.21	-	0.0
PROPKA	0.91	0.57	0.69
PypKa	1.03	0.59	0.58
DeepKa[b]	1.03 (0.87 w/o Y/C)	0.51 (0.59 w/o Y/C)	0.23 (0.33 w/o Y/C)
PKAI+	1.01	0.54	0.27
pKALM	1.07 1.25[c]	0.53 0.47	0.29 0.21
GearNet[d]	1.35 ± 0.01 (1.33)	0.42 ± 0.01 (0.45)	0.22 ± 0.00 (0.24)
ESM-2[d]	1.29 ± 0.01 (1.23)	0.22 ± 0.00 (0.25)	0.15 ± 0.00 (0.18)
GSnet (rSASA-opt)[e]	1.01 ± 0.00 (0.97)	0.56 ± 0.00 (0.60)	0.28 ± 0.00 (0.38)
GSnet (random)	1.35 ± 0.00 (1.28)	0.43 ± 0.01 (0.52)	0.19 ± 0.00 (0.23)
aLCnet (opt-AMBER)[f]	0.81 ± 0.00 (0.78)	0.75 ± 0.00 (0.76)	0.55 ± 0.01 (0.67)
aLCnet (opt-CHARMM)[f]	0.95 ± 0.00 (0.89)	0.62 ± 0.00 (0.68)	0.38 ± 0.01 (0.59)
aLCnet (random)	1.16 ± 0.01 (1.02)	0.39 ± 0.02 (0.56)	0.04 ± 0.01 (0.19)

[a]For GearNet, ESM-2, and our models, 20 training runs were conducted and the top 10 models based on validation RMSE were selected for evaluation on the test set; performance results are given as mean ± SEM with the best (out of 10) result given in parentheses. "random" indicates that a model was trained from random initial weights. Predictions with DeepKa were obtained via the DeepKa web server.[68] Predictions with PypKa,[71] PKAI+,[70] and pKALM[48] were done with software downloaded from the respective GitHub archives described in the publications. All predictions were made using experimental PDB structures. [b]Predictions for tyrosine and cysteine residues obtained via null model, metrics excluding these residues shown in parentheses. [c]For a subset of 102 out or 143 data points without redundancy to pKALM training set.[48] [d]Embeddings passed to the same MLP architecture as in our models. [e]Pretrained with six original physicochemical properties + residue-level SASA and optimized. [f]Pretrained with CpHMD pK_a data from PHMD549 data set[36] and optimized

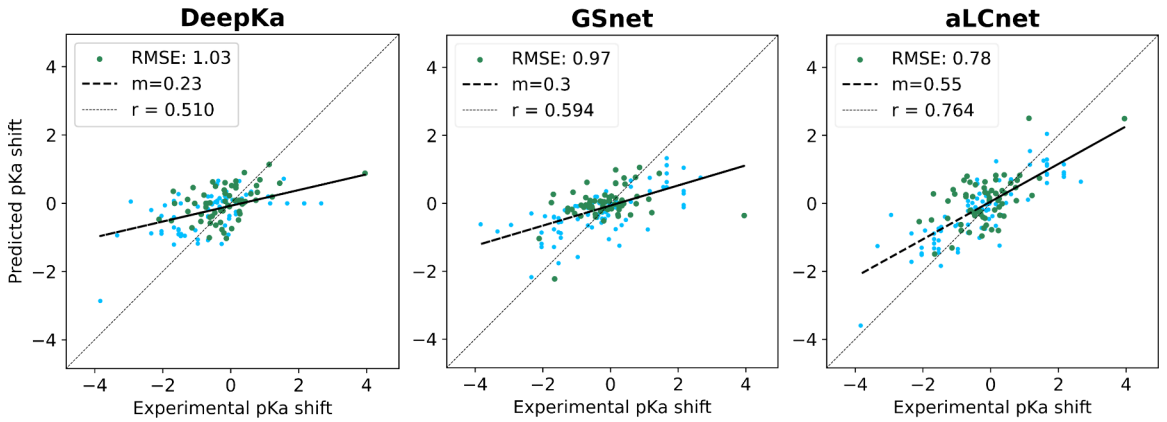


Figure 5. Experimental p$K_a$shift predictions with DeepKa, GSnet, and aLCnet on MSU-pK_a-test. For GSnet and aLCnet models the test performance is shown for the model with the lowest validation RMSE (out of 20 trained models). Performance data for DeepKa were obtained via predictions made with the DeepKa web server.[68] DeepKa predicted shifts for tyrosine and cysteine residues were obtained with the null model. Darker green points represent structures that have low fold-level similarity (TM-score <0.6) to all the structures in the training set. RMSE values are given in pK units, for all data points on the plots. Predictions were made using experimental PDB structures.

validation performance was achieved with a 5 Å edge cutoff and a maximum of 150 edges per node (Tables S5 and S6). This model was trained directly on the PHMD549 data to make pK_a predictions and achieved significantly better performance than the GSnet-based models (Table 2). The average RMSE of aLCnet on the EXP67S set was 1.10, with the best model achieving an RMSE of 1.08, exceeding the performance of PROPKA (RMSE = 1.12). However, aLCnet still could not reach the reported accuracy of DeepKa (RMSE = 0.97) or CpHMD simulations (RMSE = 1.01). The performance of GSnet and aLCnet can be further compared by looking at the distribution of individual data points (Figure S20). GSnet-based predictions are generally more conservative, closer to the null model, with only few predictions giving shifts of more than 2 pK_a units, and the slope of predicted vs experimental pK_a values is only about 0.3. With aLCnet, larger shifts are predicted, giving an improved slope of 0.48, similar to DeepKa, but with larger deviations from the experimental values. The significant improvement with aLCnet over GSnet was likely due to the inclusion of atomic charge as an input feature, as well as its truly atomic resolution, highlighting the limitations of using a generic structure-based embedding, such as GSnet trained on global molecular properties, for predictions requiring higher resolution and additional features.

Accurate Predictions of Experimental pK_a Values Using Embeddings. Training based on calculated pK_a shifts in the section above resulted in good accuracy for predicting experimental pK_a values, especially with aLCnet, but the lack of experimental data in the training set likely limited the performance. On the other hand, training a machine learning model only on experimental pK_a data is problematic because there are only relatively few data points available, especially when eliminating redundant data for the same residue in the same or similar protein structures.[44,45] Attempting to overcome these challenges, we pursued two distinct transfer learning strategies with GSnet (see Figure S6) and aLCnet (see Figure

https://doi.org/10.1021/acs.jctc.4c01682
J. Chem. Theory Comput. 2025, 21, 4830−4845

S7) embeddings that were fine-tuned along with the output MLP by training on experimental pK_a shifts.

In order to train using experimental data, we constructed the MSU-pK_a-training, MSU-pK_a-validation, and MSU-pK_a-test data sets, primarily from entries in PKAD-1[44] and PKAD-2[45] as described in the Methods section. The separation between training, validation, and test data was done carefully to avoid redundancy and to prevent data leakage from training to test performance. We note that overlap between training and test data likely resulted in overly optimiztic performance estimates in other studies.[35]

Following the transfer learning strategies, we achieved an average RMSE of 1.01 pK_a units with GSnet and an average of 0.95 using aLCnet embeddings with CHARMM-derived charges, better than values obtained with the null model[69] (1.21), slightly better than values from the machine learning methods DeepKa[68] (1.03), PKAI+[70] (1.01), pKALM[48] (1.07), better than the physics-based predictor PypKa[71] (1.03) and only slightly worse than PROPKA[72] (0.91) (Table 3). However, when aLCnet was trained using charges derived from the AMBER force field, the RMSE improved further, to 0.81, outperforming PROPKA and suggesting that AMBER-derived charges provide a more accurate representation of electrostatic interactions in our model. On the test set, aLCnet predictions with AMBER charges achieved a slope of about 0.6 when comparing predicted shifts to experimental shifts (Figure 5) whereas DeepKa, GSnet, and CHARMM-based aLCnet predictions had shallower distributions due to underpredicting larger shifts (Table 3, Figure 5). Results stratified by maximum TM-score to any training set protein show no clear performance differences between test proteins with maximum TM-scores above or below 0.6, indicating that model predictions are not strongly dependent on fold-level similarity. For aLCnet, we tested various graph construction cutoffs (6, 8, 10, 12, 14, and 16 Å) and found that 10 Å yielded the best performance (Table S7).

We note that DeepKa does not predict pK_a shifts for tyrosine and cysteine residues, so performance was evaluated both by applying the null model instead and by excluding tyrosine and cysteine residues (Table 3). To further assess the impact of this limitation, we ran DeepKa, as well as our best GSnet and aLCnet models, on the subset of MSU-pK_a-test that excluded tyrosine and cysteine residues. We found that RMSE and correlation improved for all three models when these nonionizable residues were removed (Figure S21). Notably, aLCnet still achieved the best RMSE (0.78) and correlation (0.76), but DeepKa (RMSE = 0.87, r = 0.59) outperformed GSnet (RMSE = 0.91, r = 0.49) (Figure S21). This suggests that GSnet's apparently superior performance on the full MSU-pK_a-test data set may have been influenced by the application of the null model to tyrosine and cysteine residues when evaluating DeepKa.

Because the MSU-pK_a-test data set was more concentrated around the mean than the EXP67s test set used by Cai et al.,[36] we also constructed a modified version of MSU-pK_a-test by systematically removing data points closest to the mean until its standard deviation was within 0.1 pK_a units of EXP67s (Figure S22). We then evaluated DeepKa, GSnet, and aLCnet on this subset, and we found higher RMSE and higher correlation for all 3 models than with the full MSU-pK_a-test set. aLCnet continued to achieve the best RMSE (0.84) and highest correlation (0.78), followed by GSnet (RMSE = 1.06, r = 0.61) and DeepKa (RMSE = 1.11, r = 0.56) (Figure S23). Slopes of linear regression fits to the data were largely unaffected (Figure S23). A similar pattern was observed when data points corresponding to tyrosine and cysteine residues were removed from this subset as with the full MSU-pK_a-test set. Again, aLCnet maintained the lowest RMSE (0.80) and highest correlation (0.75), while DeepKa (RMSE = 0.95, r = 0.65) outperformed GSnet (RMSE = 1.01, r = 0.51) (Figure S24). This further supports the idea that GSnet's seemingly better performance relative to DeepKa in the full MSU-pK_a-test data set may have been influenced by the application of the null model to tyrosine and cysteine residues.

Overall, the relative performance between aLCnet and the other two models remained consistent across different subsets of the test data, suggesting that the pretrained aLCnet provides more reliable pK_a predictions than GSnet and DeepKa, independent of concentration of data around the mean, and independent of inclusion or exclusion of the nonionizable residues.

A direct comparison with other machine learning methods[35,73] was not done because of significant overlap between the training sets used in the other methods and the MSU-pK_a-test or because code was not available as for the recently published predictors KaML-CBtree and KaML-GAT.[47] However, a similar RMSE value of 1.00 as with our model was reported with the P-SPOC model when testing on sequences that were not included during training.[73] Using pretrained models resulted in significantly better performance than training from random initial weights. For the GSnet architecture we achieved an RMSE of 1.35 without pretraining, compared to 1.01 with pretraining. For aLCnet we found an RMSE of 1.16 without pretraining, compared to 0.95 with pretraining. These results highlight the significant advantage of transfer learning when training with sparse data.

We also evaluated performance of the models by residue type, and these results are provided in Table S8. For aspartate residues, PROPKA and aLCnet performed similarly well with RMSEs near 0.9, with GSnet performing worse than the null model. All models, including the null model, performed similarly well for glutamate, with an RMSE around 0.6, except aLCnet with an RMSE of about 0.7. GSnet performed exceptionally well for histidine (RMSE = 0.73), with PROPKA (RMSE = 1.22) and aLCnet (1.15) performing worse. For lysine residues, aLCnet, PROPKA, and the null model achieved RMSEs near 0.5, while GSnet achieved RMSEs near 0.95. GSnet and aLCnet outperformed the null model (RMSE = 1.72) for tyrosine residues with RMSEs of 1.11 and 1.08, respectively, but they were slightly worse than PROPKA (RMSE = 0.82). GSnet performed exceptionally well for cysteine residues with an RMSE of 0.11, and aLCnet was able to exceed the performance of the null model (RMSE = 1.65) as well with an RMSE of 1.03. PROPKA performed quite poorly for cysteine with an RMSE of 4.10.

We tested whether alternative pretrained embedding models, namely GearNet (structure-based) and ESM-2 (sequence-based), could perform similarly well as GSnet and aLCnet. With GearNet we achieved an average RMSE of 1.35 while ESM-2 embeddings appeared to work slightly better, with an average RMSE of 1.29. We note that our results using ESM-2 embeddings are consistent with the performance of pKALM,[48] which also uses ESM-2 embeddings, when applied to the subset of our test cases that are nonredundant to the training set of pKALM[48] according to our similarity classification strategy. pKALM (RMSE = 1.25, r = 0.47) seems to perform slightly better than our ESM-2 model (RMSE = 1.29, r = 0.22), which may be due to the use of a BiLSTM in pKALM, as opposed to the simple MLP in our model. Either way, we speculate that that

https://doi.org/10.1021/acs.jctc.4c01682
J. Chem. Theory Comput. 2025, 21, 4830−4845

the significantly better test performance reported in the pKALM paper is due to data leakage between training and test sets. Based on our results, it seems that neither GearNet nor ESM-2 embeddings allow for accurate pK_a predictions, as neither model, even when selecting the best of all trained models, was able to exceed the performance of the null model (RMSE = 1.21).

Predictions of Experimental pK_a Values in IDPs. We further tested the performance of GSnet and aLCnet-based pK_a predictions values by applying these models to α-synuclein, an IDP. Predictions were compared with experimental pK_a measurements for 18 glutamate residues, 6 aspartate residues, and 1 histidine residue.[74] Because the structure of an IDP is inadequately described by a single structure,[75] we made predictions over an ensemble of 300 simulated structures and averaged the predictions for each residue. We found that the averaged pK_a value predictions were generally in good agreement with the experimental values, with RMSE values of 0.20 and 0.24 for GSnet and AMBER-based aLCnet predictors, respectively, across all 25 residues for which experimental pK_a values are available (Figure 6). We also tested the CHARMM-based aLCnet predictor and observed a systematic underestimation of most pK_a values relative to experimental measurements (Figure S25). This underestimation persisted across different selection radii, with no clear improvement at larger radii (Figure S25). The origin of the systematic differences in the pK_a predictions for α-synuclein using AMBER- or CHARMM-derived charges is unclear and may be subject to further investigation in future work.

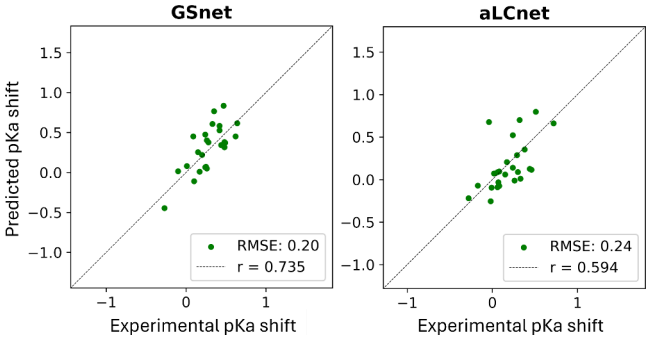


Figure 6. GSnet and aLCnet predictions of p$K_a$shifts for selected residues in α-synuclein. Experimental values were obtained via NMR in 150 mM NaCl by Croke et al..[74] RMSE values are given in pK units. Predictions were made over an ensemble of 300 simulated structures and averaged.

High-Throughput Predictions. Once trained, machine learning frameworks are very fast, especially when making multiple predictions, as they can be made in parallel on GPU computing hardware. Thus, a significant advantage of GSnet, and the related aLCnet, is vastly reduced computational cost. For instance, when making global property predictions for 4-aminobutyrate aminotransferase (UNIPROT ID P80404), a 500-residue protein, it takes more than 4 min to produce results with HYDROPRO and more than 45 min to obtain solvation free energies with APBS, whereas the trained GSnet model only requires about 1 s to make a forward pass. (Table S9). Apart from faster predictions for 4-aminobutyrate aminotransferase, GSnet also uses much less memory, about 670 MB, compared to 1.7 GB with HYDROPRO and 30.8 GB with APBS (Table S9). The fast calculation of electrostatic solvation free energies is especially interesting in the application of continuum solvation models, whose applications have been plagued by the relatively high cost of obtaining solvation free energies using traditional approaches.[76] However, further efforts are needed to increase the accuracy of solvation free energy predictions for practical applications, which is beyond the scope of the present work.

When considering pK_a predictions, it took about 180 s with GSnet and 130 s with aLCnet to predict 2000 residues on one NVIDIA GeForce RTX 2080 Ti GPU, and the cost with aLCnet includes the time to perform hydrogen bond optimization and to generate input charges via PDB 2PQR, which can likely be optimized in a future version. For comparison, predictions with PROPKA take about 231 s for 2000 residues. This opens up high-throughput pK_a predictions for very large complexes or for a large number of structures. However, we acknowledge that further validation with experimental data or high-accuracy physics-based models such as constant pH molecular dynamics may be needed to confirm the accuracy of the predictions. Either way, to illustrate a possible application, we calculated pK_a shifts for all ionizable residues in the GroEL-GroES chaperonin complex (Figure 7). The resulting shifts mapped onto the complex structure show distinct patterns of pK_a shifts, for example significant shifts of basic residues near the conical top toward acidic pH values (Figure 7).

CONCLUSIONS

The work presented here demonstrates the utility of GNNs like GSnet and aLCnet in predicting physicochemical properties of proteins while simultaneously generating molecular embeddings that can be employed in the prediction of other properties via transfer learning. We show that using embeddings and transfer learning results in better model performance, especially when training on new properties for which only sparse data is available. We demonstrate successful transfer learning to predict solvent-accessible surface areas, and we make pK_a shift predictions at accuracies that are competitive with methods developed specifically for such applications. Moreover, using GSnet and aLCnet embeddings, we achieve better accuracy than using previously developed embeddings, namely ESM-2 and GearNet.

We find good performance using GSnet, a global structure embedding, but the best performance for pK_a predictions required the higher resolution, charge-aware aLCnet. In principle, it should be possible to learn atomistic features together with global structural properties, and future work may focus on generating a more comprehensive embedding that better bridges different scales. Presumably, such an embedding would involve a higher-capacity network and would need to be trained on a larger variety of training data. For example, one could imagine training such an embedding on local electrostatic properties or properties that capture local solvation environments such as Poisson−Boltzmann-dervied Generalized Born radii.[78] Additionally, future work may benefit from incorporating conformational dynamics, as experimental properties constitute ensemble averages over thermally fluctuating molecules, rather than the static structures considered here.

As for the accurate prediction of pK_a values, we find very good performance based on aLCnet embeddings that matches the accuracy of other empirical methods but may not yet surpass the accuracy of simulation-based constant-pH approaches.[10,79] The main limitation is likely the limited amount of experimental data available for training, especially when considering the need for nonredundant data for developing transferable models that perform well beyond the training set. Using computational data

https://doi.org/10.1021/acs.jctc.4c01682
J. Chem. Theory Comput. 2025, 21, 4830−4845

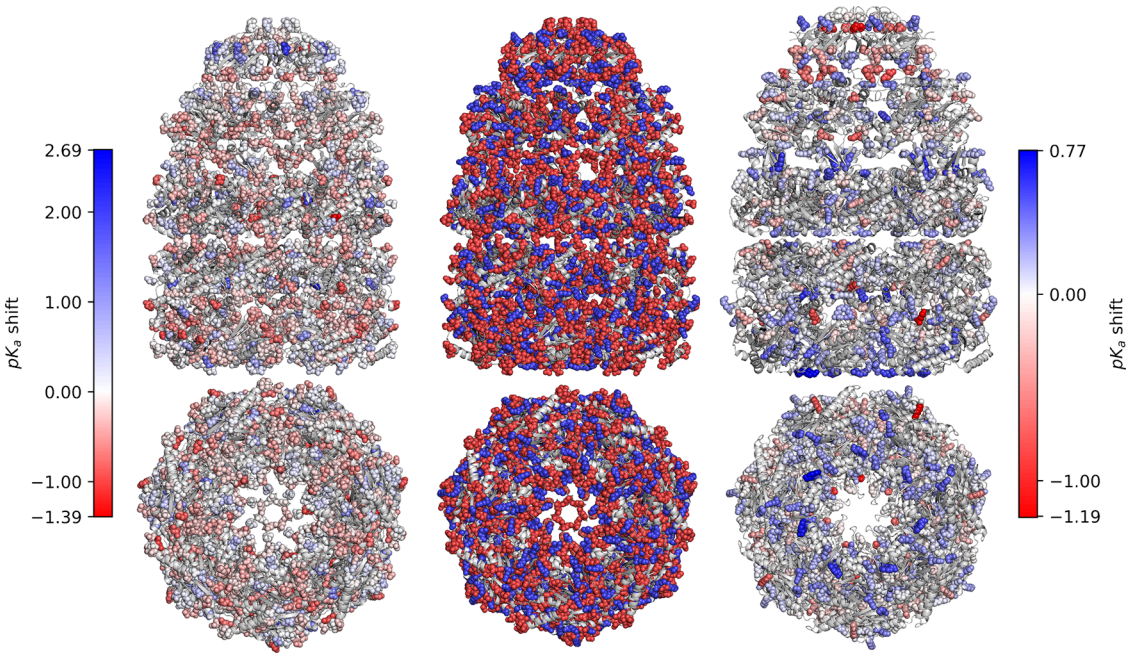


Figure 7. Predicted pK_a shifts for GroEL-GroES by aLCnet. Predicted shifts of acidic and basic residues are shown on the left and right, respectively. The color scale indicates the magnitude of the shift, with deeper shades of blue indicating a greater increase in pK_a, deeper shades of red indicating a greater decrease in pK_a, and white indicating no shift. The central panel indicates all ionizable residues, with acidic residues colored red and basic residues colored blue. The figure was generated using PyMol.[77] Predictions were made for the experimental PDB structure 1AON.

for training appears to partially resolve the issue as demonstrated previously,[36,68] but to make further progress with ML-based pK_a predictors, much more extensive experimental data is probably needed.

On the practical level, our aLCnet-based pK_a predictor is computationally very efficient, offering accuracy similar to PROPKA but at greater speed. This opens up new applications where pK_a shift predictions could be integrated into structural analysis pipelines. It is also becoming possible now to apply pK_a shift predictions to very large complexes and to large numbers of structures up to proteome-wide analyses.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.4c01682.

Supporting Tables S1-S9, Figures S1-S25, and supplementary references related to those tables and figures are provided as supporting material (PDF).

■ AUTHOR INFORMATION

Corresponding Author

Michael Feig − *Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0001-9380-6422; Phone: +1-517-432-7439; Email: mfeiglab@gmail.com

Authors

Spencer Wozniak − *Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, United States*

Giacomo Janson − *Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan 48824, United States;* orcid.org/0000-0003-1757-4193

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.4c01682

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Lim Heo for his valuable insights regarding ESM-2 and his help in generating the IDP data. We also acknowledge discussions with Alexander Jussupow and Gilberto Valdes-Garcia. Funding was provided by the National Institute of Health (NIGMS) grant R35 GM126948 and by the National Science Foundation, grant MCB 2210228.

■ REFERENCES

(1) Axen, S. D.; Huang, X.-P.; Cáceres, E. L.; Gendelev, L.; Roth, B. L.; Keiser, M. J. A Simple Representation of Three-Dimensional Molecular Structure. *J. Med. Chem.* **2017**, *60*, 7393−7409.

(2) Faulon, J.-L.; Bender, A. *Handbook of chemoinformatics algorithms*; CRC press, 2010.

(3) Isert, C.; Atz, K.; Schneider, G. Structure-based drug design with geometric deep learning. *Curr. Op. Struct. Biol.* **2023**, *79*, No. 102548.

(4) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: consistent treatment of internal and surface residues in empirical p K a predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525−537.

(5) Jurrus, E.; Engel, D.; Star, K.; Monson, K.; Brandi, J.; Felberg, L. E.; Brookes, D. H.; Wilson, L.; Chen, J.; Liles, K.; Chun, M.; Li, P.; Gohara, D. W.; Dolinsky, T.; Konecny, R.; Koes, D. R.; Nielsen, J. E.; Head-Gordon, T.; Geng, W.; Krasny, R.; Wei, G.; Holst, M. J.; McCammon, J. A.; Baker, N. A. Improvements to the APBS biomolecular solvation software suite. *Protein Sci.* **2018**, *27*, 112−128.

(6) de la Torre, J. G.; Huertas, M. L.; Carrasco, B. Calculation of hydrodynamic properties of globular proteins from their atomic-level structure. *Biophys. J.* **2000**, *78*, 719−730.

(7) King, E.; Aitchison, E.; Li, H.; Luo, R. Recent developments in free energy calculations for drug discovery. *Front. Mol. Biosci.* **2021**, *8*, No. 712085.

(8) Bueschbell, B.; Caniceiro, A. B.; Suzano, P. M.; Machuqueiro, M.; Rosário-Ferreira, N.; Moreira, I. S. Network biology and artificial intelligence drive the understanding of the multidrug resistance phenotype in cancer. *Drug Resistance Updates* **2022**, *60*, No. 100811.
(9) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037−10041.
(10) de Oliveira, V. M.; Liu, R.; Shen, J. Constant ph molecular dynamics simulations: Current status and recent applications. *Curr. Opin. Struct. Biol.* **2022**, *77*, No. 102498.
(11) Qing, R.; Hao, S.; Smorodina, E.; Jin, D.; Zalevsky, A.; Zhang, S. Protein design: From the aspect of water solubility and stability. *Chem. Rev.* **2022**, *122*, 14085−14179.
(12) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.
(13) Zhang, Z.; Xu, M.; Jamasb, A.; Chenthamarakshan, V.; Lozano, A.; Das, P.; Tang, J. Protein representation learning by geometric structure pretraining. *arXiv* preprint **2022**, arXiv:2203.06125 DOI: .
(14) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513−530.
(15) Fabian, B.; Edlich, T.; Gaspar, H.; Segler, M.; Meyers, J.; Fiscato, M.; Ahmed, M. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv* preprint **2020**, arXiv:2011.13230 DOI: .
(16) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J. Cheminf.* **2021**, *13*, 1−23.
(17) Peteani, G.; Huynh, M. T. D.; Gerebtzoff, G.; Rodríguez-Pérez, R. Application of machine learning models for property prediction to targeted protein degraders. *Nat. Commun.* **2024**, *15*, 5764.
(18) Deng, J.; Yang, Z.; Wang, H.; Ojima, I.; Samaras, D.; Wang, F. A systematic study of key elements underlying molecular property prediction. *Nat. Commun.* **2023**, *14*, 6395.
(19) Hosna, A.; Merry, E.; Gyalmo, J.; Alom, Z.; Aung, Z.; Azim, M. A. Transfer learning: a friendly introduction. *J. Big Data* **2022**, *9*, 102.
(20) Farahani, A.; Pourshojae, B.; Rasheed, K.; Arabnia, H. R. A concise review of transfer learning. In *International conference on computational science and computational intelligence (CSCI)*, 2020; IEEE: pp 344−351,
(21) Mayr, F.; Wieder, M.; Wieder, O.; Langer, T. Improving small molecule pK_a prediction using transfer learning with graph neural networks. *Front. Chem.* **2022**, *10*, No. 866585.
(22) Bepler, T.; Berger, B. Learning the protein language: Evolution, structure, and function. *Cell Syst.* **2021**, *12*, 654−669.e3.
(23) Neyshabur, B.; Sedghi, H.; Zhang, C. What is being transferred in transfer learning? *Adv. Neural Inf. Process. Sys.* **2020**, *33*, 512−523.
(24) Jensen, R. Behaviorism, latent learning, and cognitive maps: needed revisions in introductory psychology textbooks. *Behavior Analyst* **2006**, *29*, 187−209.
(25) Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R. et al.Relational inductive biases, deep learning, and graph networks. *arXiv* preprint **2018**, arXiv:1806.01261 DOI: .
(26) Gligorijević, V.; Renfrew, P. D.; Kosciolek, T.; Leman, J. K.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B. C.; Fisk, I. M.; Vlamakis, H.; Xavier, R. J.; Knight, R.; Cho, K.; Bonneau, R. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **2021**, *12*, 3168.
(27) Wieder, O.; Kohlbacher, S.; Kuenemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer, T. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today: Technol.* **2020**, *37*, 1−12.
(28) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural message passing for quantum chemistry. In *International conference on machine learning*, 2017; PMLR: pp 1263−1272.
(29) Chen, C.; Zhou, J.; Wang, F.; Liu, X.; Dou, D. Structure-aware protein self-supervised learning. *Bioinformatics* **2023**, *39*, No. btad189.
(30) Heinzinger, M.; Littmann, M.; Sillitoe, I.; Bordin, N.; Orengo, C.; Rost, B. Contrastive learning on protein embeddings enlightens midnight zone. *NAR: Genomics Bioinf.* **2022**, *4*, No. lqac043.
(31) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.
(32) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123−1130.
(33) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **2022**, *50*, D439−D444.
(34) Gokcan, H.; Isayev, O. Prediction of Protein pKa with Representation Learning. *Chem. Sci.* **2022**, *13*, 2462−2474.
(35) Chen, A. Y.; Lee, J.; Damjanovic, A.; Brooks, B. R. Protein pKa Prediction by Tree-Based Machine Learning. *J. Chem,. Theory Comp.* **2022**, *18*, 2673−2686.
(36) Cai, Z.; Liu, T.; Lin, Q.; He, J.; Lei, X.; Luo, F.; Huang, Y. Basis for Accurate Protein pK_a Prediction with Machine Learning. *J. Chem. Inf. Model.* **2023**, *63*, 2936−2947.
(37) Ortega, A.; Amorós, D.; De La Torre, J. G. Prediction of hydrodynamic and other solution properties of rigid proteins from atomic-and residue-level models. *Biophys. J.* **2011**, *101*, 892−898.
(38) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **2015**, *109*, 1528−1532.
(39) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; De Groot, B. L.; Grubmüller, H.; MacKerell, A. D., Jr CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Meth.* **2017**, *14*, 71−73.
(40) Wang, G.; Dunbrack, R. L., Jr PISCES: a protein sequence culling server. *Bioinformatics* **2003**, *19*, 1589−1591.
(41) Wang, W.; Peng, Z.; Yang, J. Single-sequence protein structure prediction using supervised transformer protein language models. *Nat. Comp. Sci.* **2022**, *2*, 804−814.
(42) Valdes-Garcia, G.; Heo, L.; Lapidus, L. J.; Feig, M. Modeling concentration-dependent phase separation processes involving peptides and RNA via residue-based coarse-graining. *J. Chem,. Theory Comp.* **2023**, *19*, 669−678.
(43) Heo, L.; Feig, M. One bead per residue can describe all-atom protein structures. *Structure* **2024**, *32*, 97−111.e6.
(44) Pahari, S.; Sun, L.; Alexov, E. PKAD: a database of experimentally measured pKa values of ionizable groups in proteins. *Database* **2019**, *2019*, baz024.
(45) Ancona, N.; Bastola, A.; Alexov, E. PKAD-2: new entries and expansion of functionalities of the database of experimentally measured pKa's of proteins. *J. Comp. Biophys. Chem.* **2023**, *22*, 515−524.
(46) Wilson, C. J.; Karttunen, M.; de Groot, B. L.; Gapsys, V. Accurately Predicting Protein pKa Values Using Nonequilibrium Alchemy. *J. Chem,. Theory Comp.* **2023**, *19*, 7833−7845.
(47) Shen, M.; Kortzak, D.; Ambrozak, S.; Bhatnagar, S.; Buchanan, I.; Liu, R.; Shen, J. KaMLs for Predicting Protein pKa Values and Ionization States: Are Trees All You Need? *bioRxiv* **2024**, 2024.2011.2009.622800. DOI: .

(48) Xu, S.; Onoda, A. Accurate and Rapid Prediction of Protein pKa: Protein Language Models Reveal the Sequence-pKa Relationship. *bioRxiv* **2024**, 2024.2009.2016.613101. DOI: .
(49) Zhang, Y.; Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005**, *33*, 2302−2309.
(50) Sievers, F.; Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **2018**, *27*, 135−145.
(51) Chakrabarti, P.; Pal, D. The interrelationships of side-chain and main-chain conformations in proteins. *Prog. Biophys. Mol. Biol.* **2001**, *76*, 1−102.
(52) Shi, Y.; Huang, Z.; Feng, S.; Zhong, H.; Wang, W.; Sun, Y. Masked label prediction: Unified message passing model for semi-supervised classification. *arXiv* preprint **2020**, arXiv:2009.03509 DOI: .
(53) Lei Ba, J.; Kiros, J. R.; Hinton, G. E. Layer normalization. *arXiv* preprint **2016**, arXiv:1607.06450 DOI: .
(54) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. Schnet−a deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, No. 241722.
(55) Satorras, V. G.; Hoogeboom, E.; Welling, M. E(n) equivariant graph neural networks. In *Proceedings of the 38th International conference on machine learning*, 2021; PMLR: pp 9323−9332.
(56) Vaswani, A.; Shazeer, N. M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.
(57) Li, H.; Zhang, X.; Liu, X.; Gong, Y.; Wang, Y.; Yang, Y.; Chen, Q.; Cheng, P. Gradient-Mask Tuning Elevates the Upper Limits of LLM Performance. *arXiv* preprint **2024**, arXiv:2406.15330 DOI: .
(58) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* **2019**, *32* 80206.
(59) Fey, M.; Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. *arXiv* preprint **2019**, arXiv:1903.02428 DOI: .
(60) Kingma, D.; Adam, P. A method for stochastic optimization. *arXiv* preprint **2014**, arXiv:1412.6980 DOI: .
(61) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(62) Brenner, S. E.; Chothia, C.; Hubbard, T. J. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 6073−6078.
(63) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L., III Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *J. Comput. Chem.* **2004**, *25*, 265−284.
(64) Zhou, Y.; Feig, M.; Wei, G.-W. Highly accurate biomolecular electrostatics in continuum dielectric environments. *J. Comput. Chem.* **2008**, *29*, 87−97.
(65) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; et al. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545−1614.
(66) Singh, A. Structure prediction for orphan proteins. *Nat. Meth.* **2023**, *20*, 176−176.
(67) Einstein, A. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Ann. Phys.* **1905**, *322*, 549−560.
(68) Cai, Z.; Peng, H.; Sun, S.; He, J.; Luo, F.; Huang, Y. DeepKa Web Server: High-Throughput Protein pKa Prediction. *J. Chem. Inf. Model.* **2024**, *64*, 2933−2940.
(69) Thurlkill, R. L.; Grimsley, G. R.; Scholtz, J. M.; Pace, C. N. pK values of the ionizable groups of proteins. *Protein Sci.* **2006**, *15*, 1214−1218.
(70) Reis, P. B. P. S.; Bertolini, M.; Montanari, F.; Rocchia, W.; Machuqueiro, M.; Clevert, D.-A. A Fast and Interpretable Deep Learning Approach for Accurate Electrostatics-Driven pKa Predictions in Proteins. *J. Chem,. Theory Comp.* **2022**, *18*, 5068−5078.
(71) Reis, P. B.; Vila-Vicosa, D.; Rocchia, W.; Machuqueiro, M. PypKa: A flexible python module for poisson−boltzmann-based p K a calculations. *J. Chem. Inf. Model.* **2020**, *60*, 4442−4448.
(72) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J. Chem., Theory Comput.* **2011**, *7*, 2284−2295.
(73) Liu, S.; Yang, Q.; Zhang, L.; Luo, S. Accurate Protein pKa Prediction with Physical Organic Chemistry Guided 3D Protein Representation. *J. Chem. Inf. Model.* **2024**, *64*, 4410−4418.
(74) Croke, R. L.; Patil, S. M.; Quevreaux, J.; Kendall, D. A.; Alexandrescu, A. T. NMR determination of pKa values in α-synuclein. *Protein Sci.* **2011**, *20*, 256−269.
(75) Oldfield, C. J.; Uversky, V. N.; Dunker, A. K.; Kurgan, L. Chapter 1 - Introduction to intrinsically disordered proteins and regions. *Intrinsically Disord. Proteins* **2019**, 1−34.
(76) Feig, M.; Chocholoušová, J.; Tanizaki, S. Extending the horizon: towards the efficient modeling of large biomolecular complexes in atomic detail. *Theor. Chem. Acc.* **2006**, *116*, 194−205.
(77) DeLano, W. L. Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr.* **2002**, *40*, 82−92.
(78) Onufriev, A.; Case, D. A.; Bashford, D. Effective Born radii in the generalized Born approximation: The importance of being perfect. *J. Comput. Chem.* **2002**, *23*, 1297−1304.
(79) Wallace, J. A.; Shen, J. K. Predicting pKa values with continuous constant pH molecular dynamics. *Methods Enzymol.* **2009**, *466*, 455−475.

https://doi.org/10.1021/acs.jctc.4c01682
J. Chem. Theory Comput. 2025, 21, 4830−4845