Representing local protein environments with atomistic foundation models

Meital Bojan [*1] **Sanketh Vedula** [*1 2 3] **Advaith Maddipatla** [1 2 4] **Nadav Bojan Sellam** [1] **Federico Napoli** [1] **Paul Schanda** [1] **Alex M. Bronstein** [1 2]

Abstract

The local structure of a protein strongly impacts its function and interactions with other molecules. Therefore, a concise, informative representation of a local protein environment is essential for modeling and designing proteins and biomolecular interactions. However, these environments' extensive structural and chemical variability makes them challenging to model, and such representations remain under-explored. In this work, we propose a novel representation for a local protein environment derived from the intermediate features of atomistic foundation models (AFMs). We demonstrate that this embedding effectively captures both local structure (e.g., secondary motifs), and chemical features (e.g., amino-acid identity and protonation state). We further show that the AFM-derived representation sp ace exhibits meaningful structure, enabling the construction of data-driven priors over the distribution of biomolecular environments. Finally, in the context of biomolecular NMR spectroscopy, we demonstrate that the proposed representations enable a first-of-its-kind physics-informed chemical shift predictor that achieves state-of-the-art accuracy. Our results demonstrate the surprising effectiveness of atomistic foundation models and their emergent representations for protein modeling beyond traditional molecular simulations. We believe this will open new lines of work in constructing effective functional representations for protein environments.

1. Introduction

Proteins are complex three-dimensional structures composed of hundreds to thousands of atoms, arranged in a specific manner in space. These molecular architectures as well as their dynamic fluctuations are crucial for the function of a protein. Local environments within a protein are highly diverse, due to the variability of the sequence of amino acids and the fold of the amino-acid chain to a 3D object. Computationally representing these inherently heterogeneous, dynamic, and chemically diverse environments is challenging, yet very important: many key properties of proteins critically rely on these local enviroments, for example the ability of a protein to bind a ligand or perform enzymatic activity. Dynamics of a protein is ultimately rooted in the packing in 3D space, and is generally considered crucial for function. In this context, *biomolecular nuclear magnetic resonance* (NMR) spectroscopy is a powerful technique, allowing to probe protein dynamics and interactions at atomic resolution.

The principal NMR observables are *chemical shifts* of certain atomic nuclei endowed with non-zero spin (e.g.,^{1}H, ^{13}C, ^{15}N atoms). The chemical shift of an atom is exquisitely sensitive to its local electronic environment and is critically dependent on local structural features, such as the type of amino acid, the dihedral angles between bonds (and, thus, secondary structure), the orientation of side chains, hydrogen bonding, and solvent exposure. These parameters, averaged over time, impact the chemical shift in a complex manner, and while the chemical shift is exquisitely sensitive to them, interpreting chemical shifts in terms of structure is hampered by this complexity.

There is a strong need for accurate computational prediction of chemical shifts, with enormous potential in several directions: accurately predicted chemical shifts from a given protein structure largely facilitates the assignment of observed NMR signals to individual atoms; moreover, the inverse direction of determining structures and dynamics from experimentally observed chemical shifts hinges on the ability to predict chemical shifts with high accuracy.

Existing chemical-shift prediction approaches (Neal et al., 2003; Shen & Bax, 2008; Zhou et al., 2023) use experimentally determined chemical shifts and experimental 3D structures; these structures are parametrized with hand-crafted descriptors to represent the local environment (e.g. dihedral angles, hydrogen bonds, ring-current shifts). Likewise, pre-

[*]Equal contribution [1]IST Austria [2]Technion, Israel [3]Princeton University, USA [4]University of Oxford, UK. Correspondence to: Meital Bojan <meital.bojan@campus.technion.ac.il>, Sanketh Vedula <svedula@princeton.edu>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

dictors of the protonation state of titratable groups equally employ hand-crafted features (Olsson et al., 2011). These tools face limitations in adaptability and accuracy. This highlights the need for more expressive, effective, and generalizable representations of local protein environments.

In computational chemistry, descriptors like Parrinello-Behler symmetry functions (Behler, 2011; Behler & Parrinello, 2007; Behler, 2016) and Smooth Overlap of Atomic Positions (SOAP) are widely used to represent molecular environments (Jäger et al., 2018). These methods encode atomic interactions and geometry into concise, invariant representations suited for geometric modeling. Modern atomistic machine learning techniques (Schütt et al., 2018; 2021) implicitly learn similar representations within neural network layers, achieving accuracy comparable to density functional theory (DFT) (Deng et al., 2023) in tasks like molecular dynamics (MD) simulations.

Recent advances have shifted neural-network interatomic potentials from element-specific models to universal atomistic foundation models[1] (AFMs) trained on large datasets of molecules with DFT-simulated energies. ANI-2x pioneered this transition, covering biologically relevant elements (H, C, N, O, F, S, Cl) with DFT-level accuracy across millions of organic molecules (Smith et al., 2018). Three recent model families – Multiplicative Atomic Convolutional Equivariant network (MACE) (Batatia et al., 2022), OrbNet (Qiao et al., 2020), and AIMNet (Zubatyuk et al., 2019) – have further advanced the field: MACE combines equivariant message passing with high-order many-body terms for materials science applications; OrbNet enhances graph neural networks with semi-empirical orbital features, reducing data requirements for DFT-level accuracy by an order of magnitude; and AIMNet-2 extends ANI with learned atomic embeddings and multitask training (energies, charges, spin), improving transferability to charged and open-shell systems.

In this study, atomistic foundation models are used to describe local protein structure. While typically applied to molecular simulations (Kapil et al., 2024; Lahey et al., 2020), our focus is different: analyzing AFM internal representations to reveal the biochemical and biophysical information they encode.

Contributions and Summary of Results In what follows, we outline our major contributions and provide a summary of results.

Canonical local environments and AFM representations. Protein environments vary widely, making their direct comparison challenging. To make residue-level representations comparable, we introduce *canonical environments*: regions centered on a *focus residue* and containing every amino acid whose atom coordinates lie within 5Å (Hausdorff distance) of a focus residue. We then construct transferable *environment representations* from the AFM embeddings of the backbone atoms and their bonded hydrogens – N, CA, CB, C, O, H, HA – whose roles are conserved across residues (see Fig. 1). To our knowledge, this is the first use of AFMs to characterize local protein structure.

AFM representations effectively capture local protein structure and chemistry. By analyzing over 165K protein environments extracted from 1327 non-redundant chains, we demonstrate that their embeddings reveal secondary-structure features and amino-acid identities (see Fig. A.6). In addition, we train lightweight task networks on these embeddings to predict protonation states obtained from a Poisson–Boltzmann solver (Reis et al., 2020), and find that they achieve higher accuracy than pKa-ANI (Gokcan & Isayev, 2022) and propKa (Olsson et al., 2011) (see Table B.3).

AFM representations allow computation of similarity metrics and calculating likelihoods. We show that AFM embeddings have a meaningful latent space. Specifically, distances between environment embeddings reflect both geometric similarity and chemical context, enabling intuitive clustering of related local motifs. We construct likelihood and conditional likelihoods in the AFM embedding space and show that they effectively capture the distribution of biomolecular environments (see Fig. A.7, A.13). We show that the defined likelihoods capture subtle structural deformations in protein structure and allow capturing distributions of shifts, making them valuable for structural quality assessment, uncertainty estimation, anomaly detection, dataset curation, and potentially in guiding Boltzmann generators (Noé et al., 2019). Furthermore, leveraging the AFM embedding space, we introduce a similarity metric to compare protein environments (see Appendix G.1).

AFMs enable physics-grounded accurate chemical shift predictor. In the context of biomolecular NMR, we leverage AFM embeddings to predict chemical shifts in proteins. We demonstrate that a lightweight downstream network receiving these embeddings as the inputs outperforms the current state-of-the-art chemical shift predictor, UCBShift2-X, on backbone and side-chain atoms (see Fig. 3). We conduct a case study examining the ring current effects produced by aromatic side chains on chemical shifts of neighboring nuclei, and demonstrate that our predictor follows physically consistent trends, whereas UCBShift2-X shows unphysical behavior (see Fig. 2). We also pair the predictor with AFM-derived likelihoods to assign confidence scores, yielding a reliable uncertainty measure around the predicted shifts (see Fig. A.13).

[1]We use "atomistic foundation models" to denote transferable force fields and universal neural-network potentials. Analogous to foundation models in vision and language, they learn compact representations of atomistic systems that can be reused for a variety of downstream tasks as we demonstrate here.

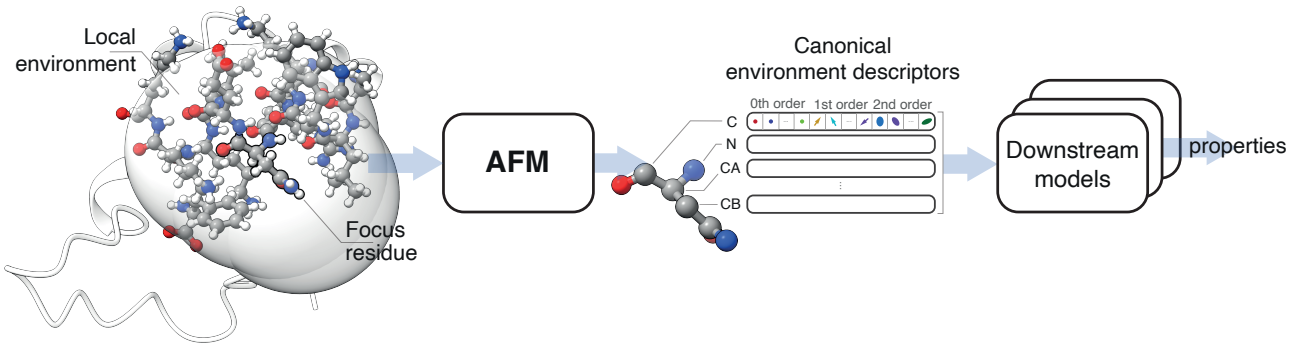

Figure 1: **Proposed construction of local protein environment descriptors and their use.** A focus residue is used as the center of an environment including full amino acids located in the 5Å range from it. The atoms are embedded by an atomistic foundational model (AFM), producing a descriptor for each atom in the environment influenced by all the other atoms. The embeddings corresponding to the principal atoms of the focus residue form a canonical descriptor of the environment and may comprise zeroth and higher-order irreducible representation components. These descriptors are then used by downstream models predicting atom, residue, or environment properties.

Interpreting AFM representations. We probe the content of AFM embeddings in two controlled settings. Firstly, we systematically modify side chain orientations and track how the embeddings respond (see Fig. 2 and A.5). Secondly, we follow an unfolding simulation in which a seven-residue α-helix extends into a strand (see Fig. A.8, A.12). In both cases, principal-component analysis uncovers clear latent directions that mirror the underlying structural changes. We then ask whether an environment descriptor alone suffices to reconstruct the local structure. By optimizing AlphaFold3 (Abramson et al., 2024) structures with only this descriptor, we obtain alternative conformations that reproduce the backbone arrangement and many side-chain orientations, although fine-grained details remain imperfect (Maddipatla et al., 2025; 2024; Levy et al., 2025). This partial recovery suggests that AFM embeddings encode most—but not all—of the information required to rebuild a protein environment (see Appendix. I.1).

Benchmarking different AFMs. We benchmark three model families—MACE & EGRET, OrbNet, and AIMNet—on tasks defined on protein environment that include secondary-structure assignment, amino-acid identification, acid dissociation constant (pK_a) and chemical-shift prediction (see Tables B.2, B.1, B.4). MACE-based embeddings perform best on every task except pK_a prediction, where AIMNet exhibits superior performance (see Table B.3).

These findings show that AFM embeddings compactly and effectively describe complex biomolecular environments, we believe this would inspire future studies of local protein structure with AFMs. The rest of the paper is organized as follows: In Section 2 we provide a brief introduction to AFMs and dedicate Section 3 to the construction of the proposed local environment descriptors. In Section 4, we demonstrate that the proposed representations capture local protein structure and chemistry. In Section 5, we discuss the evaluation of likelihood and similarity of protein environments. Section 6 is dedicated to the prediction of NMR chemical shifts from AFM representations. In Section 7, we discuss the interpretation of the proposed representations. Section 8 briefly overviews related work and Section 9 discusses limitations and future research directions. We encourage the reader to have a detailed look at the main and supplementary figures visualizing the main findings of the paper.

2. Background on Atomistic Foundation Models

Force field models approximate a system's potential energy from atomic positions and types. The gradient of this energy with respect to atomic positions defines the forces acting on each atom. Traditional force fields were modelled by empirically computing interatomic potentials or using DFTs to analytically determine interatomic forces (Finnis, 2003). More recently, AFMs, have been developed to approximate the energy function and force fields from computational quantum mechanical data. In this work, we use the following AFMs (detailed summary in Table B.5 in Appendix):

MACE (Batatia et al., 2022; Kovács et al., 2023) is based on the body-ordered Atomic Cluster Expansion (ACE) (Dusson et al., 2022) with higher-order (many-body) equivariant message passing to capture complex atomic interactions. It constructs atomic energies using symmetry-preserving basis functions and updates them through rotation-equivariant neural interactions. MACE is trained on datasets containing small molecules and tripeptides to predict the energies and forces of DFT relaxation trajectories. Introduced recently,

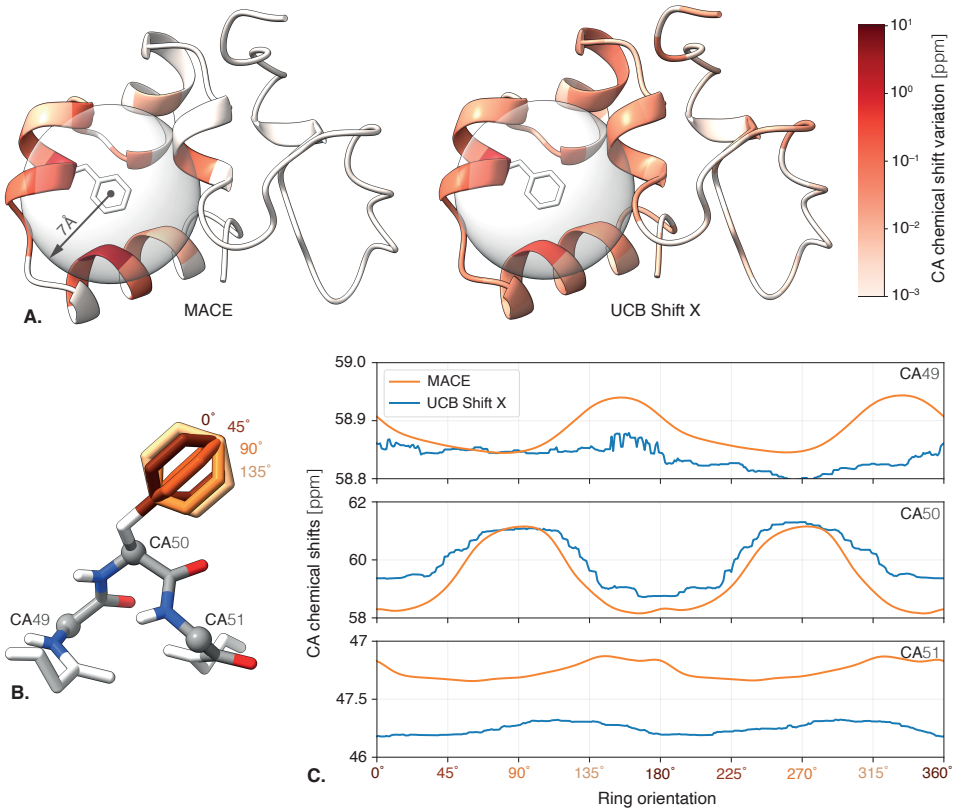

Figure 2: **Synthetic example showing the influence of a phenylalanine sidechain aromatic ring on surrounding chemical shifts.** (A) The magnitude of change of backbone CA chemical shifts over different ring orientations as predicted using the proposed MACE-based predictor (left) and UCB Shift X (right). A 7Å sphere indicates the radius from the ring center at which the influence of the ring current is expected to become negligible. Note that UCB Shift X predicts much longer-range, albeit small, ring influence extending beyond 20Å. (B) Locations of three nearby CA atoms; and (C) their predicted chemical shifts plotted vs. the ring orientation. Note the smooth 180°-periodic behavior of the Mace shift prediction and the decay of the effect scale with the distance from the ring.

Egret (Wagen et al., 2025) shares a similar higher-order equivariant message-passing architecture with MACE, but introduces architectural optimizations that improve computational efficiency and scalability.

OrbNet (Qiao et al., 2020) enhances graph neural networks by incorporating quantum-derived features calculated using fast approximate quantum chemistry methods. These features include symmetry-adapted atomic orbital overlaps and elements of the Fock matrix (Szabo & Ostlund, 1996), which capture important information about the electronic structure. OrbNet is trained on DFT relaxation trajectories and quantum chemistry datasets (containing small molecules and materials) to predict molecular energies and properties.

AIMNet (Zubatyuk et al., 2019) uses message passing over atomic environment descriptors and iterative updates inspired by charge equilibration schemes. Through these updates, AIMNet captures both short- and long-range interactions and builds a joint, information-rich representation of each atom in its molecular context. The model is trained on diverse quantum mechanical datasets to predict molecular energies, forces, and other properties across diverse chemical systems.

3. Representing protein local environments with AFMs

Notation. Denote a residue in a protein as a, and the all-atom representation of the protein structure as $\mathcal{X} = \{(\mathbf{x}_1, z_1) \dots (\mathbf{x}_m, z_m)\}$, where $\mathbf{x}_j \in \mathbb{R}^3$ and z_j are the location and atomic number of jth atom in $\mathcal{X}$, respectively. Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ be the AFM with parameters θ, and $\mathcal{Y} = \{(\mathbf{y}_1, z_1), \dots, (\mathbf{y}_m, z_m)\}$ the atom-wise feature representations of the AFM, where $\mathbf{y} \in \mathbb{R}^d$. The *local environment* of a *focus residue* a is denoted by $\mathcal{X}_a \subseteq \mathcal{X}$. The subset of atoms in residue a used for predicting biochemical properties is denoted as $\mathcal{A}_a$. Given a subset of atom indices $\mathcal{A}_a$, we denote by $\mathcal{Y}_{\mathcal{A}_a} = \{(\mathbf{y}_i, z_i) \mid i \in \mathcal{A}_a\} \subseteq \mathcal{Y}$ the restriction of $\mathcal{Y}$ to those atoms.

We seek a representation of a local protein environment to

be (i) sensitive to local changes, (ii) insensitive to global variations, (iii) fast to compute, (iv) canonical, enabling direct comparison across diverse environments, and (v) effective and generalizable to unseen environments. Encoding full proteins with several thousand atoms with AFMs is computationally inefficient and redundant since we are solely interested in locally sensitive features. To construct such representations, we propose to construct a local environment around a focus residue to be encoded later by the AFM.

Creating local environments. To balance computational efficiency while retaining local structural context, given a focus residue a, we construct the environment $\mathcal{X}_a$ as the union of all residues whose atoms are at most 5Å away from the atoms of a. The procedure is described in Algorithm 1.

AFM representations. Given an environment $\mathcal{X}_a$, AFMs produce atom-level features over the layers of the network. These learned atomistic features are *contextual*, influenced not only by the atom itself but also by the other atoms in the environment due to the message-passing operations within AFMs. Given $\mathcal{X}_a$, atom-wise feature representations are extracted from the final layer of the AFM. Each AFM produces representations of shape $N \times d$, where $N = |\mathcal{X}_a|$ and d is the dimension of the representation. In order to construct representations that are *canonical*, i.e., comparable across residues, the representations corresponding to atoms in $\mathcal{A}$ are retained, yielding contextualized representations of their local chemical environments per residue. Given a protein structure, this process is performed for each residue in the sequence, and residue-wise features are curated. A visual depiction of this procedure is presented in Fig. 1.

4. AFM representations capture local protein structure and chemistry

In what follows, we investigate how AFM representations of local environments can be leveraged to predict a range of structural and chemical properties.

Amino acid identity & secondary structure. A protein's *primary structure* is a linear sequence of amino acids connected by covalent peptide bonds. The twenty standard proteinogenic amino acids (Kawashima & Kanehisa, 2000) share a backbone of four heavy atoms: the α-carbon (CA), amide nitrogen (N), carbonyl carbon (C), and carbonyl oxygen (O), with variable side chains (R-group). A protein's *secondary structure* is a local motif of backbone conformation, stabilized by hydrogen bonds between atoms of non-adjacent amino acids (Kabsch & Sander, 1983), with the most common types being α-helices (H) and β-strands (E).

To examine the information encoded by the AFM representations, the atomistic features from MACE were projected into a two dimensional space using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). The resulting embeddings, shown in Fig. A.6, are annotated with amino acid labels, secondary structure, and dihedral angles. Notably, features corresponding to α-helices and β-strands form distinct clusters in the UMAP space. A similar pattern of separation is observed when embeddings are annotated by amino acid identity and backbone dihedral angles, indicating that chemically and structurally distinct features are well-represented in the feature space.

To quantitatively assess this capability, we train classifiers to predict the secondary structures and the amino acid identities. Further details on the loss function, target label space, and the definition of $\mathcal{A}_a$ can be found in Appendix F.1, while a comprehensive description of the model architecture is provided in Appendix F.3. Classifier accuracy was evaluated using precision, recall, and F1 scores. Results for secondary structure prediction are summarized in Table B.1, while amino acid classification results are presented in Table B.2. For secondary structure, models trained with MACE and Egret features consistently achieve superior prediction on average. For amino acid prediction, models trained using Egret features consistently outperform those trained with alternative representations.

Acid dissociation constants ($\mathbf{pK_a}$). pK_a quantifies the tendency of a titratable group (such as the carboxylic acid group of an Asp side chain) to donate a proton, serving as a measure of its acidity. When the solvent pH is lower than pK_a, the molecule will be primarily in its protonated form, while being increasingly deprotonated at pH higher than pK_a. The pK_a value is the pH value at which the group is half protonated over the ensemble of molecules. Although each amino acid has a canonical pK_a value in isolation, its actual value within a protein can deviate due to variations in the local chemical environment. External factors such as (i) the presence of nearby charged residues, polar atoms, or functional groups, and (ii) limited solvent accessibility—particularly in buried regions—can influence protonation behavior. This makes pK_a prediction an ideal test case for evaluating whether atomistic features can encode meaningful local chemical information.

For quantitative assessment, a regression model was trained on amino acids with ionizable, proton-transferring side chains to predict their pK_a values. Ground-truth pK_a values were estimated using a Poisson–Boltzmann solver (Reis et al., 2020). The evaluation focused on glutamic acid (GLU) and aspartic acid (ASP), which tend to deprotonate, and lysine (LYS) and histidine (HIS), which tend to protonate. Additional training details are provided in Appendix F. As shown in Table B.3, models trained with AIMNet features outperform those using alternative features because it is trained to predict multiple molecular properties simultaneously. All AFM-based regressors outperform conventional

baselines such as PropKA (Olsson et al., 2011) and pKa-ANI (Gokcan & Isayev, 2022).

5. Computing likelihoods and measuring similarity of protein environments

The AFM embedding space encodes local biomolecular structure, capturing structural and chemical context. We leverage the richness of AFM embedding spaces to define a distribution over biomolecular environments. Given a set of reference environments $\mathcal{E}_{\text{ref}} = \{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$ and a reference atom set $\mathcal{A}$, we define the likelihood of an environment $\mathcal{X}_a$ at a focus residue a as

$$p(\mathcal{X}_a) = \frac{1}{|\mathcal{E}_{\text{ref}}|} \sum_{\mathcal{X}_{a'} \in \mathcal{E}_{\text{ref}}} \exp\left(-\frac{\| f_\theta(\mathcal{X}_a)|_{\mathcal{Y}_\mathcal{A}} - f_\theta(\mathcal{X}_{a'})|_{\mathcal{Y}_\mathcal{A}} \|^2}{2\sigma^2} \right) \tag{1}$$

where σ refers to the *bandwidth*, and it controls the influence of each reference environment on $\mathcal{X}_a$. This is equivalent to performing kernel-density estimation in the AFM embedding space, with a radial basis function kernel. Intuitively, the likelihood measures how typical $\mathcal{X}_a$ is among the reference environments $\mathcal{E}_{\text{ref}}$. Subsequently, the conditional likelihoods can be defined similarly to Eq. 1 by curating an appropriate set of reference environments that satisfy the chosen conditioning.

Measuring distribution shifts. To evaluate the quality of unconditional likelihood estimation and its ability to detect subtle distribution shifts in the local structure, we curated a set of reference environments randomly sampling $10,000$ environments from $1,100$ protein structures that were relaxed using the Amber99 force field (Hornak et al., 2006). We then measured the likelihoods of each environment sampled from a test set of 225 proteins, before and after performing Amber99 relaxation. The results shown in Fig. A.13 demonstrate that the likelihood function is sensitive to subtle conformational changes. Relaxed structures consistently receive higher likelihoods, and the distribution of the paired differences captures fine-grained structural variations. This makes the approach well-suited for detecting out-of-distribution conformations and assessing local structural quality.

Secondary-structure conditioned likelihoods. To further evaluate the sensitivity of conditional likelihoods, we randomly sampled 1000 environments from secondary structures annotated as α-helix (H), β-strand (E), and turn (T) by DSSP, and constructed conditional likelihoods $p(\mathcal{X}|\text{H})$, $p(\mathcal{X}|\text{E})$, and $p(\mathcal{X}|\text{T})$, respectively. Fig. A.7 depicts the projections of conditional likelihood triplets $(\log p(\mathcal{X}_i|\text{E}), \log p(\mathcal{X}_i|\text{H}), \log p(\mathcal{X}_i|\text{T}))$ of $8,163$ environments from the test set. The results highlight a clear separation by true secondary structure of the environment. The one-dimensional marginal distributions further emphasize the distinct statistical profiles of each structural class.

We further direct the reader to Fig. A.10 and Appendix G.1 for a detailed analysis of the use of AFM embeddings to evaluate chemical environment similarity.

6. Physics-grounded chemical shift prediction for proteins

Background and prior art. We refer an uninitiated reader to Appendix C for a short background on chemical shifts and biomolecular NMR. Several computational methods have been developed to predict NMR chemical shifts from molecular structures (Shen & Bax, 2008; Neal et al., 2003; Han et al., 2011; Li & Brüschweiler, 2012; 2015; Kohlhoff et al., 2009; Meiler, 2003; Moon & Case, 2007), with recent approaches increasingly relying on machine learning. The current state-of-the-art, UCBShift (Zhou et al., 2023), combines sequence and structure alignment with a random forest model, using reference proteins and structural descriptors to guide predictions. Its latest iteration, UCBShift 2.0 (Ptaszek et al., 2024), extends this framework to include side-chain atoms alongside backbone predictions. Despite their effectiveness, these methods depend on reference-based similarity measures, which limits their generalizability.

AFM-based shift predictor. Chemical shift prediction is formulated as a downstream task over AFM embeddings. A lightweight graph neural network takes AFM representations as input and predicts the chemical shifts of a target atom. Separate models are trained for the N, CA, C, H, and HA backbone atoms, as well as the CB, CG, CD, CD2, CG2, and CE side-chain atoms. Dataset curation, model architecture, and training details are provided in Appendix D, F.3.

As evident from Fig. 3 , the proposed predictor outperforms UCBShift2 for both backbone and side-chain heavy atoms with the exception of the hydrogen HA. A comprehensive quantitative comparison of chemical shift prediction accuracy across different AFM variants is provided in Table. B.4. We observe that MACE-based models consistently achieve the best performance.

To further probe the shift predictor and evaluate whether it correctly captures the local structural influence on the shift, we carefully design the following three case studies.

Effects of ring currents. An external magnetic field induces a flow of delocalized π electrons in an aromatic ring, which in turn produces its own magnetic field. Because chemical shifts depend on the net magnetic field produced by the local structural environment, they are strongly influenced by such ring currents. As the phenylalanine ring has a C_2 symmetry, states separated by rotation of the ring by 180° are indistinguishable. To probe whether our shift

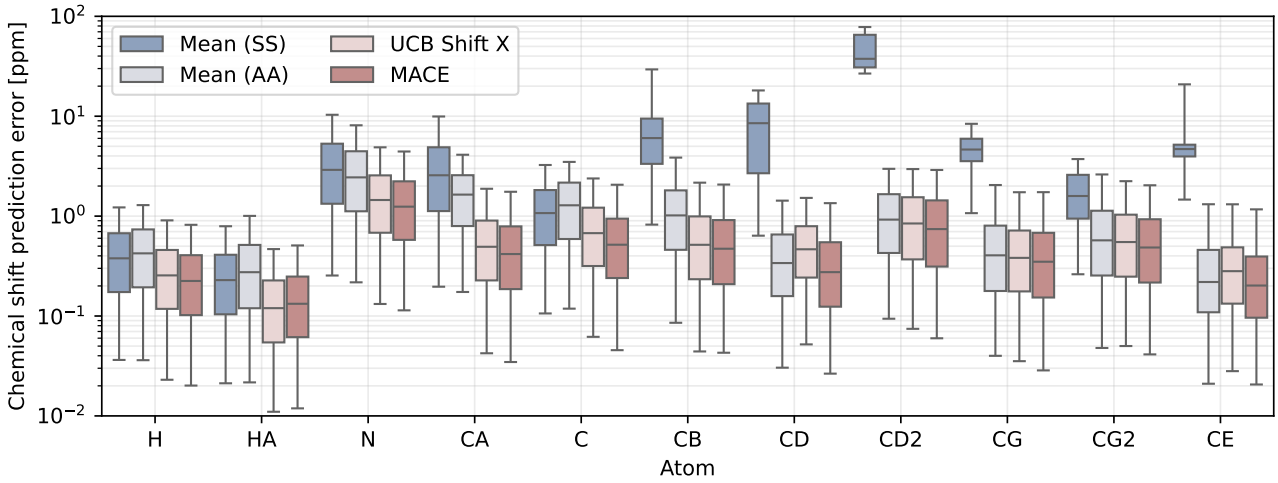

Figure 3: **Chemical shift prediction errors for different atom types** evaluated on a test set of $132,228$ environments from 203 non-redundant BMRB records with experimentally determined chemical shifts used as the reference. The median prediction error in ppm and the $25\% - 75\%$ (boxes) and $5\% - 95\%$ (whiskers) confidence intervals are depicted.

predictor captures ring-current effects accurately, we synthetically modified the side chain of a Phe (residue 50 in PDB ID: `1ZV6`) by rotating its aromatic ring ring about the χ_2 dihedral angle from $-180°$ to $180°$. Fig. 2 depicts the studied environments and the predicted chemical shifts. Our AFM-based shift predictor shows the expected $180°$ periodicity (Haigh & Mallion, 1979) in the chemical shifts of backbone CA atoms in the vicinity of the ring, with the ring current influence decaying smoothly as the distance from the ring increases and becoming negligible after 7Å. In contrast, UCBShift extends this influence beyond the expected range and fails to reproduce the smoothness and periodicity expected in theory.

Helix unfolding into a strand. Chemical shifts of CA and CB have distinct behaviors in helices and strands. We perform an MD simulation of a 7-residue peptide unfolding from a helix to a strand. Predicted chemical shifts along the trajectory are depicted in Fig. A.12. We note that our method correctly reproduces the experimentally observed chemical shift distributions for CA and CB atoms. It clearly captures the characteristic pattern where CA shifts decrease in strands compared to helices, while CB shifts exhibit the opposite trend.

Alternative conformations. We evaluate our shift predictor over a 100 ns MD simulation of the protein `4OLE:B` from (Rosenberg et al., 2024), which features two stable conformers: conformer A containing a helix and conformer B with a linearly structured loop (Fig. A.9). Our method captures distinct secondary chemical shift distributions for lysine 63 CA and CB atoms. The helical conformer exhibits the expected increase in CA secondary shift and decrease in CB secondary shift.

Next, we investigate whether the aforementioned likelihood computation for an environment can reliably anticipate the resulting chemical shift prediction error.

Uncertainty estimation. We first build the unconditional likelihood model $p(\mathcal{X})$ for biomolecular environments on the training set used for shift prediction. We then evaluate this likelihood for every test environment and record the corresponding chemical shift prediction error. Fig. 4 reports the distribution of errors, stratified by likelihood. Environments with lower likelihoods consistently exhibit larger prediction errors, indicating that the likelihood can serve as a practical measure of *confidence* in chemical shift prediction tasks. To the best of our knowledge, ours is the first chemical shift predictor to supply such a confidence score.

We summarize by concluding that AFM-based shift prediction not only delivers better accuracy, but is also physics-grounded and allows us to provide confidence estimates for each prediction.

7. Interpreting AFM representations

While atomistic features from AFMs excel at representing local protein structure and chemistry, it remains unclear how their internal representations respond to structural perturbations or whether they encode meaningful physical properties. To address this, the behavior of AFM representations under specific conformational changes is examined. Such changes induce smooth, interpretable trajectories in the embedding space, suggesting that the features capture physical motions and respect constraints such as locality, continuity, and symmetry. These properties are explored in the following case studies.

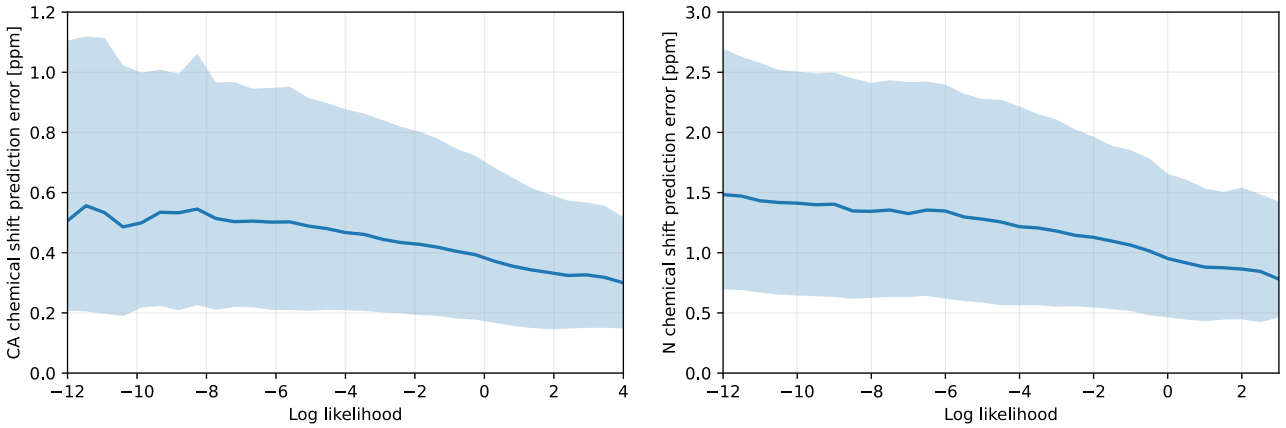

Figure 4: Chemical shift prediction accuracy of CA (left) and N (right) atoms stratified by the KDE-estimated likelihood of the corresponding MACE descriptors. Depicted are the median and 25%-50% confidence intervals. Higher-likelihood environments correspond to lower prediction error and can be used as an uncertainty measure.

Rotated phenylalanine sidechain aromatic ring. Using the synthetically modified phenylalanine described above, we perform PCA on the MACE embeddings obtained at every χ_2 angle for the nearby CA atoms (residues 50 and 51 of PDB ID: `1ZV6`). We observe that the first two principal components trace a smooth, one-dimensional curve, capturing the expected $180°$ periodicity of the rotation. However, the effect diminishes with distance: residue 49 begins to deviate from the periodic pattern, while distant residue 40 exhibits no discernible structure, illustrating the spatial locality of the descriptor's sensitivity.

Helix unfolding. In context of MD simulations of helix-to-strand transition described earlier (Fig. A.8), we perform PCA on the MACE embeddings of CA atoms in residues 6 and 7, and obseve that they follow smooth, non-intersecting trajectories. These trajectories reflect the gradual unfolding process, indicating that the descriptor space encodes continuous and interpretable structural transitions.

Further exploration of reconstructing protein environments from AFM embeddings is presented in Appendix. I.1.

8. Related Work

Recent work has highlighted the versatility of pretrained AFMs as learned representations for molecular tasks. (Shiota et al., 2024) use MACE and M3GNet features to predict chemical shifts in small molecules, while (Elijošius et al., 2024) employ MACE embeddings for zero-shot molecular generation via evolutionary search. Unlike these small-molecule applications, our method is specifically designed for proteins — large, structurally diverse biomolecules requiring environment-specific representations to capture their complex local chemistry. Similar to our idea, (Gokcan & Isayev, 2022) predict protein pK_a values using ANI-2x AFM representations; as shown in Table B.3, our approach outperforms theirs in pK_a prediction accuracy.

9. Discussion

In this work, we introduced a new representation for local environments in proteins, and demonstrated its effectiveness across applications. While our downstream models used frozen AFM embeddings, task-specific fine-tuning of AFMs is a promising direction that may yield further gains. Moreover, our chemical shift prediction for HA, while accurate, still slightly lags behind UCBShift, pointing to opportunities for future improvement. Notwithstanding these limitations, we believe that the proposed fully differentiable chemical shift predictor can be used to guide AlphaFold and similar generative models (Maddipatla et al., 2025) for determining structure from experimentally measured chemical shifts – an important task in protein NMR that has so far been hindered by the complexity of structure-to-chemical-shift relation.

Impact Statement

The work presented here attempts to advance protein representation, since proteins are fundamental ingredients of life and are implicated in health and disease, the potential impact on basic and applied research may be profound. We do not see any special ethical concerns worth highlighting.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630 (8016):493–500, 2024.

Ahdritz, G., Bouatta, N., Floristean, C., Kadyan, S., Xia, Q., Gerecke, W., O'Donnell, T. J., Berenberg, D., Fisk, I., Zanichelli, N., et al. Openfold: Retraining alphafold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature Methods*, 21(8):1514–1524, 2024.

Batatia, I., Kovacs, D. P., Simm, G., Ortner, C., and Csányi, G. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in neural information processing systems*, 35: 11423–11436, 2022.

Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics*, 134(7), 2011.

Behler, J. Perspective: Machine learning potentials for atomistic simulations. *The Journal of chemical physics*, 145(17), 2016.

Behler, J. and Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14):146401, 2007.

Bermel, W., Bertini, I., Felli, I. C., and Pierattelli, R. Nmr spectroscopy for the structural characterization of natural products. *Natural Product Reports*, 32(4):356–367, 2015.

Boyarski, A. and Bronstein, A. Multidimensional scaling. In *Computer Vision: A Reference Guide*, pp. 836–849. Springer, 2021.

Case, D. A. Chemical shift prediction in biomolecular nmr. *Current Opinion in Structural Biology*, 23(2):172–176, 2013.

Cheng, L., Welborn, M., Christensen, A. S., and Miller, T. F. A universal density matrix functional from molecular orbital-based machine learning: Transferability across organic molecules. *The Journal of chemical physics*, 150 (13), 2019.

Chmiela, S., Vassilev-Galindo, V., Unke, O. T., Kabylda, A., Sauceda, H. E., Tkatchenko, A., and Müller, K.-R. Accurate global machine learning force fields for molecules with hundreds of atoms. *Science Advances*, 9 (2):eadf0873, 2023.

Claridge, T. D. *High-Resolution NMR Techniques in Organic Chemistry*. Elsevier, 3rd edition, 2016.

Deng, B., Zhong, P., Jun, K., Riebesell, J., Han, K., Bartel, C. J., and Ceder, G. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Dusson, G., Bachmayr, M., Csányi, G., Drautz, R., Etter, S., van Der Oord, C., and Ortner, C. Atomic cluster expansion: Completeness, efficiency and stability. *Journal of Computational Physics*, 454:110946, 2022.

Elfwing, S., Uchibe, E., and Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

Elijošius, R., Zills, F., Batatia, I., Norwood, S. W., Kovács, D. P., Holm, C., and Csányi, G. Zero shot molecular generation via similarity kernels. *arXiv preprint arXiv:2402.08708*, 2024.

Fazekas, Z., K. Menyhárd, D., and Perczel, A. Locohd: a metric for comparing local environments of proteins. *Nature Communications*, 15(1):4029, 2024.

Finnis, M. *Interatomic forces in condensed matter*, volume 1. Oxford Series on Materials Mod, 2003.

Frishman, D. and Argos, P. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, 23(4):566–579, 1995.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

Gokcan, H. and Isayev, O. Prediction of protein p k a with representation learning. *Chemical Science*, 13(8):2462–2474, 2022.

Günther, H. *NMR Spectroscopy: Basic Principles, Concepts, and Applications in Chemistry*. Wiley, 3rd edition, 2013.

Haigh, C. and Mallion, R. Ring current theories in nuclear magnetic resonance. *Progress in nuclear magnetic resonance spectroscopy*, 13(4):303–344, 1979.

Han, B., Liu, Y., Ginzinger, S., and Wishart, D. Shiftx2: significantly improved protein chemical shift prediction. *Journal of Biomolecular NMR*, 50(1):43–57, 2011. doi: 10.1007/s10858-011-9478-4.

Hoch, J. C., Baskaran, K., Burr, H., Chin, J., Eghbalnia, H. R., Fujiwara, T., Gryk, M. R., Iwata, T., Kojima, C., Kurisu, G., et al. Biological magnetic resonance data bank. *Nucleic acids research*, 51(D1):D368–D376, 2023.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725, 2006.

Jäger, M. O., Morooka, E. V., Federici Canova, F., Himanen, L., and Foster, A. S. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Computational Materials*, 4(1):37, 2018.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Kabsch, W. and Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

Kapil, V., Kovács, D. P., Csányi, G., and Michaelides, A. First-principles spectroscopy of aqueous interfaces using machine-learned electronic and quantum nuclear effects. *Faraday Discussions*, 249:50–68, 2024.

Kawashima, S. and Kanehisa, M. Aaindex: amino acid index database. *Nucleic acids research*, 28(1):374–374, 2000.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Kohlhoff, K. J., Robustelli, P., Cavalli, A., Salvatella, X., and Vendruscolo, M. Fast and accurate predictions of protein nmr chemical shifts from interatomic distances. *Journal of the American Chemical Society*, 131(39):13894–13895, 2009.

Kovács, D. P., Moore, J. H., Browning, N. J., Batatia, I., Horton, J. T., Kapil, V., Witt, W. C., Magdău, I.-B., Cole, D. J., and Csányi, G. Mace-off23: Transferable machine learning force fields for organic molecules. *arXiv preprint arXiv:2312.15211*, 2023.

Kuprov, I., Wagner-Rundell, N., and Hore, P. J. Nmr of large proteins: Overcoming the relaxation hurdle. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 50(1):1–19, 2007.

Lahey, S.-L. J., Thien Phuc, T. N., and Rowley, C. N. Benchmarking force field and the ani neural network potentials for the torsional potential energy surface of biaryl drug fragments. *Journal of Chemical Information and Modeling*, 60(12):6258–6268, 2020.

Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., et al. Drugbank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(D1):D1091–D1097, 2014.

Levy, A., Chan, E. R., Fridovich-Keil, S., Poitevin, F., Zhong, E. D., and Wetzstein, G. Solving inverse problems in protein space using diffusion-based priors, 2025. URL https://arxiv.org/abs/2406.04239.

Li, D. and Brüschweiler, R. Ppm_one: a static protein structure based chemical shift predictor. *J. Biomol. NMR*, 62(3):403–409, 2015.

Li, D.-W. and Brüschweiler, R. PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. *J. Biomol. NMR*, 54 (3):257–265, nov 2012.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Maddipatla, A., Sellam, N. B., Bojan, M., Vedula, S., Schanda, P., Marx, A., and Bronstein, A. M. Inverse problems with experiment-guided alphafold, 2025. URL https://arxiv.org/abs/2502.09372.

Maddipatla, S. A., Sellam, N. B., Vedula, S., Marx, A., and Bronstein, A. Generative modeling of protein ensembles guided by crystallographic electron densities, 2024. URL https://arxiv.org/abs/2412.13223.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Meiler, J. Proshift: protein chemical shift prediction using artificial neural networks. *Journal of Biomolecular NMR*, 26:25–37, 2003.

Moon, S. and Case, D. A. A new model for chemical shifts of amide hydrogens in proteins. *Journal of biomolecular NMR*, 38:139–150, 2007.

Neal, S., Nip, A. M., Zhang, H., and Wishart, D. S. Shiftx: chemical shift prediction from protein structures. *Journal of Biomolecular NMR*, 26(3):215–240, 2003.

Noé, F., Olsson, S., Köhler, J., and Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Olsson, M. H., Søndergaard, C. R., Rostkowski, M., and Jensen, J. H. Propka3: consistent treatment of internal and surface residues in empirical p k a predictions. *Journal of chemical theory and computation*, 7(2):525–537, 2011.

Ptaszek, A. L., Li, J., Konrat, R., Platzer, G., and Head-Gordon, T. Ucbshift 2.0: Bridging the gap from backbone to side chain protein chemical shift prediction for protein structures. *Journal of the American Chemical Society*, 146(46):19234–19245, 2024. doi: 10.1021/jacs.4c10474.

Qiao, Z., Welborn, M., Anandkumar, A., Manby, F. R., and Miller, T. F. Orbnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *The Journal of chemical physics*, 153(12), 2020.

Ramakrishnan, R., Dral, P. O., Rupp, M., and Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Reis, P. B., Vila-Vicosa, D., Rocchia, W., and Machuqueiro, M. Pypka: A flexible python module for poisson–boltzmann-based p k a calculations. *Journal of chemical information and modeling*, 60(10):4442–4448, 2020.

Rosenberg, A. A., Marx, A., and Bronstein, A. M. Codon-specific ramachandran plots show amino acid backbone conformation depends on identity of the translated codon. *Nature Communications*, 13(1):2815, 2022. doi: 10.1038/s41467-022-30390-9. URL https://doi.org/10.1038/s41467-022-30390-9.

Rosenberg, A. A., Vedula, S., Bronstein, A. M., and Marx, A. Seeing double: Molecular dynamics simulations reveal the stability of certain alternate protein conformations in crystal structures. *bioRxiv*, 2024. doi: 10.1101/2024.08.31.610605. URL https://www.biorxiv.org/content/early/2024/08/31/2024.08.31.610605.

Ruddigkeit, L., Van Deursen, R., Blum, L. C., and Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

Schütt, K., Kindermans, P.-J., Sauceda Felix, H. E., Chmiela, S., Tkatchenko, A., and Müller, K.-R. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

Schütt, K., Unke, O., and Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, pp. 9377–9388. PMLR, 2021.

Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A., and Müller, K.-R. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24), 2018.

Shen, Y. and Bax, A. Sparta+: a modest improvement in empirical nmr chemical shift prediction by means of an artificial neural network. *Journal of Biomolecular NMR*, 40(4):353–362, 2008.

Shiota, T., Ishihara, K., and Mizukami, W. Universal neural network potentials as descriptors: towards scalable chemical property prediction using quantum and classical computers. *Digital Discovery*, 3(9):1714–1728, 2024.

Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O., and Roitberg, A. E. Less is more: Sampling chemical space with active learning. *The Journal of chemical physics*, 148 (24), 2018.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Szabo, A. and Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Dover Publications, Mineola, New York, 1996. ISBN 978-0486691862.

Wagen, C. C., Mann, E. L., Vandezande, J. E., Wagen, A. M., and Schneider, S. C. Egret-1: Pretrained neural network potentials for efficient and accurate bioorganic simulation. *arXiv preprint arXiv:2504.20955*, 2025.

Wishart, D. S. Advances in nmr-based metabolomics: applications in drug discovery and development. *Current Opinion in Drug Discovery & Development*, 14(1):25–33, 2011.

Zhang, H., Neal, S., and Wishart, D. S. Refdb: a database of uniformly referenced protein chemical shifts. *Journal of biomolecular NMR*, 25:173–195, 2003.

Zhou, Y., Wang, Y., Li, H., Zhang, Y., and Xu, D. Ucbshift: Protein chemical shift prediction using transfer learning and machine learning. https://github.com/THGLab/CSpred, 2023. Accessed: 2025-05-13.

Zubatyuk, R., Smith, J. S., Leszczynski, J., and Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Science advances*, 5(8):eaav6490, 2019.

A. Additional Figures

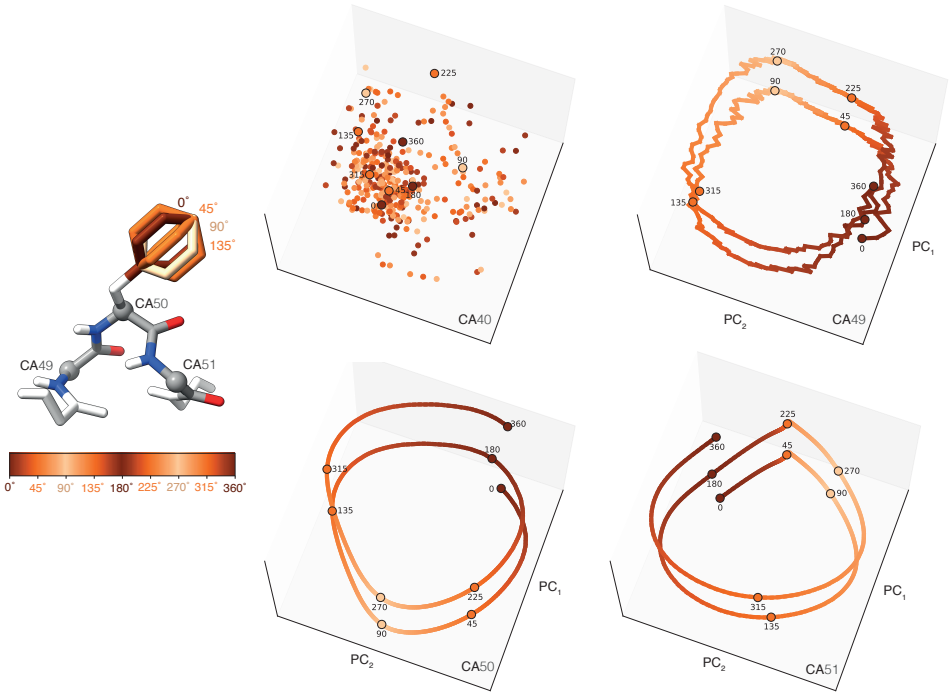

Figure A.5: **Structure of the MACE embedding space for the rotating phenylalanine aromatic ring from Fig. 2**. Left: a fragment of a protein with simulated rotation of the phenylalanine sidechain ring. Right: Mace embeddings of backbone CA atoms projected onto the first two principal components. The vertical axis separates the different rotation angles for visual clarity. Note that residues close to the ring (50 and 51) exhibit a one-dimensional structure with nearly perfect 180-degree periodicity. The farther residue 49 shows some breakdown of this structure, while residue 40, uninfluenced by the ring current, manifests a lack of relation between ring rotation and the embedding space structure.

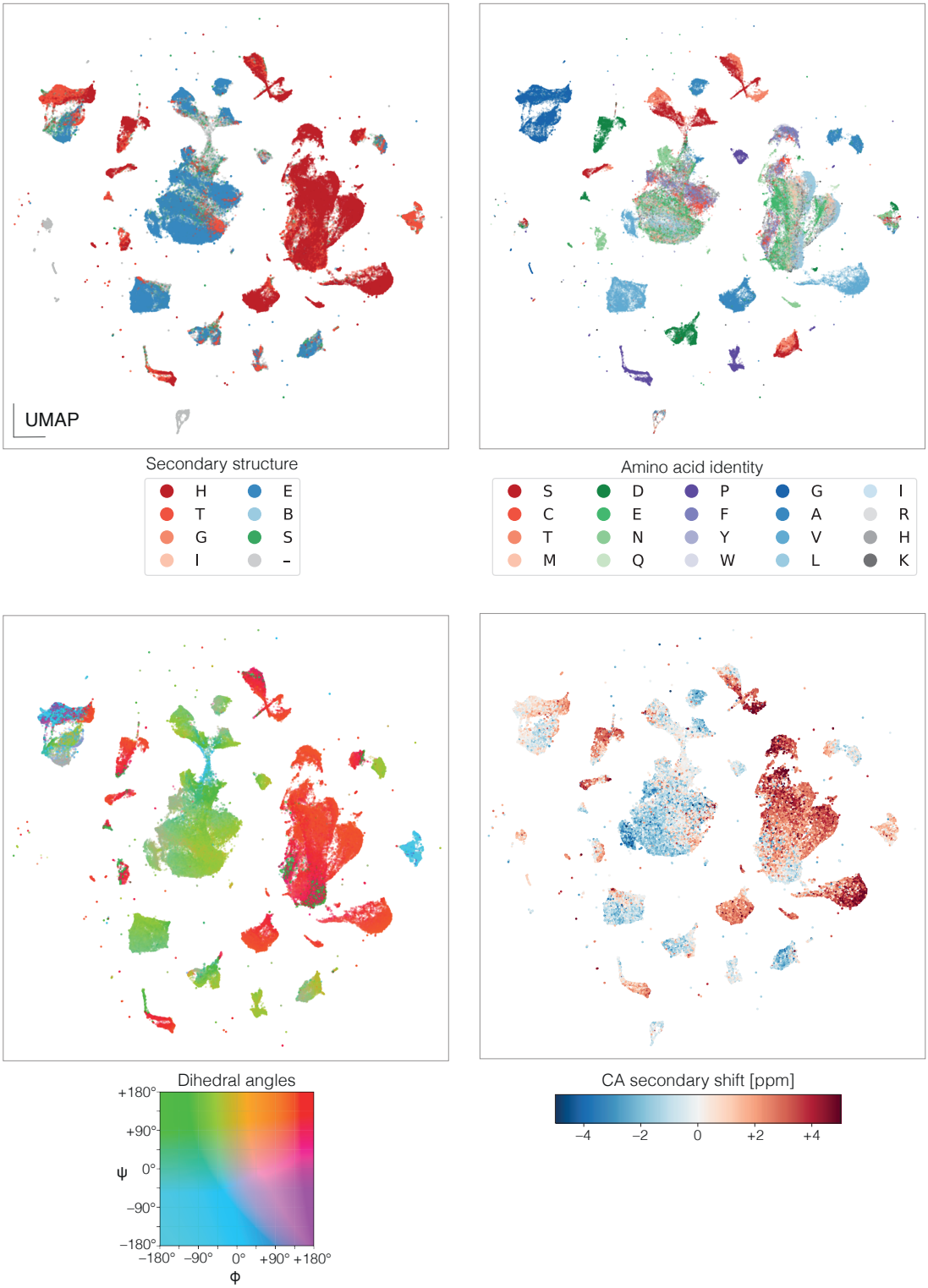

Figure A.6: **MACE embedding space reveals meaningful structural and chemical information.** Depicted are two-dimensional UMAP coordinates of 165, 913 protein environments from 1327 non-redundant chains predicted by AlphaFold2 (Jumper et al., 2021), labeled left-to-right, top-to-bottom according to the DSSP secondary structure class (Table B.6), amino acid chemical identity, the pair of backbone dihedral angles (ϕ, ψ), and CA secondary chemical shift (relative to a random coil).

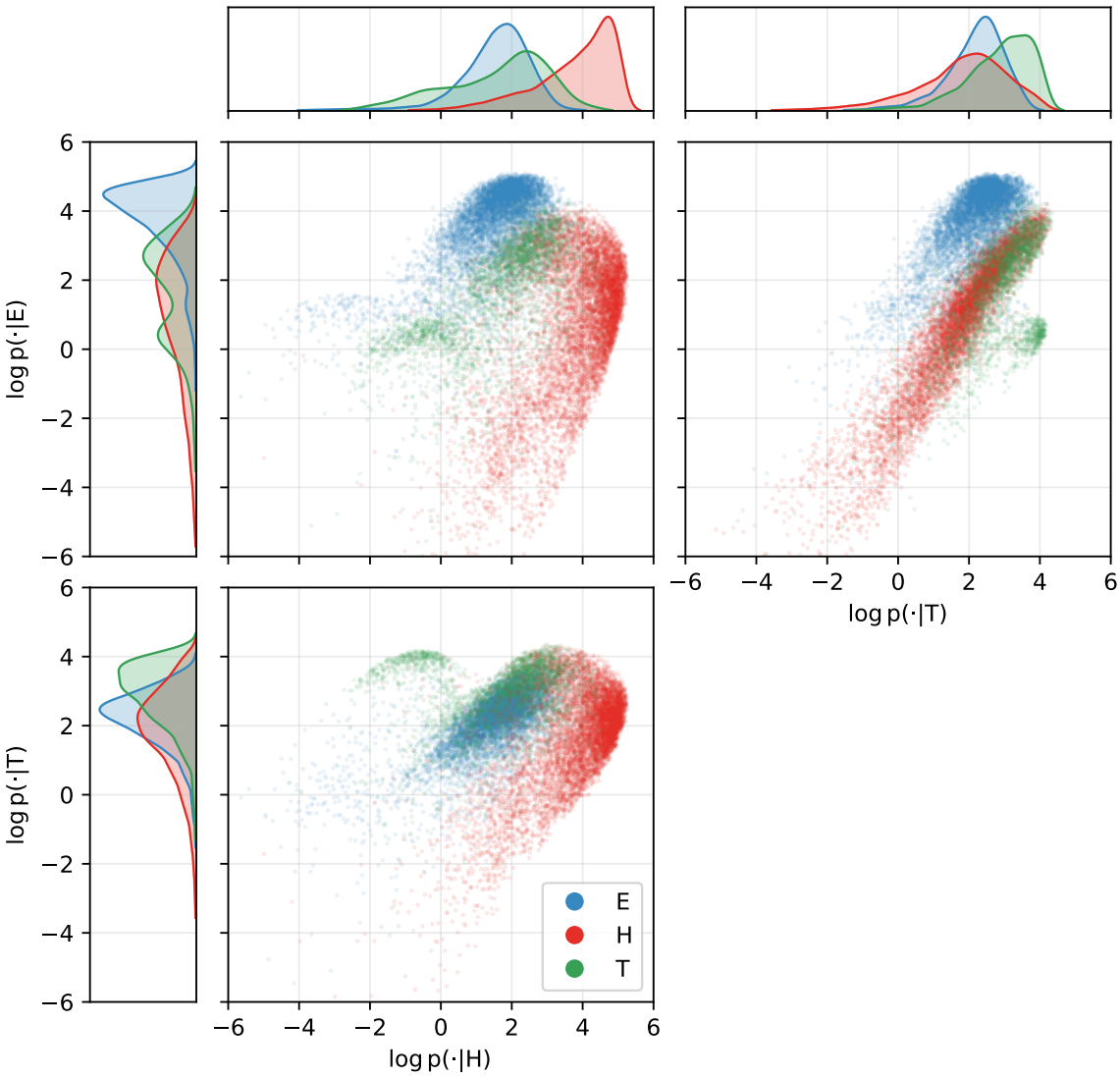

Figure A.7: **Estimated likelihoods of Mace embeddings of backbone CA atoms in different secondary structures.** The two-dimensional plots show projections of the log-likelihoods $(\log p(\mathbf{x}_i|\mathrm{E}), \log p(\mathbf{x}_i|\mathrm{H}), \log p(\mathbf{x}_i|\mathrm{T})$ of 8163 Mace embeddings $\mathbf{x}_i$ in strands (E), helices (H), and turns (T). The one-dimensional plots depict the KDE-estimated marginal distributions of each of the log-likelihoods.

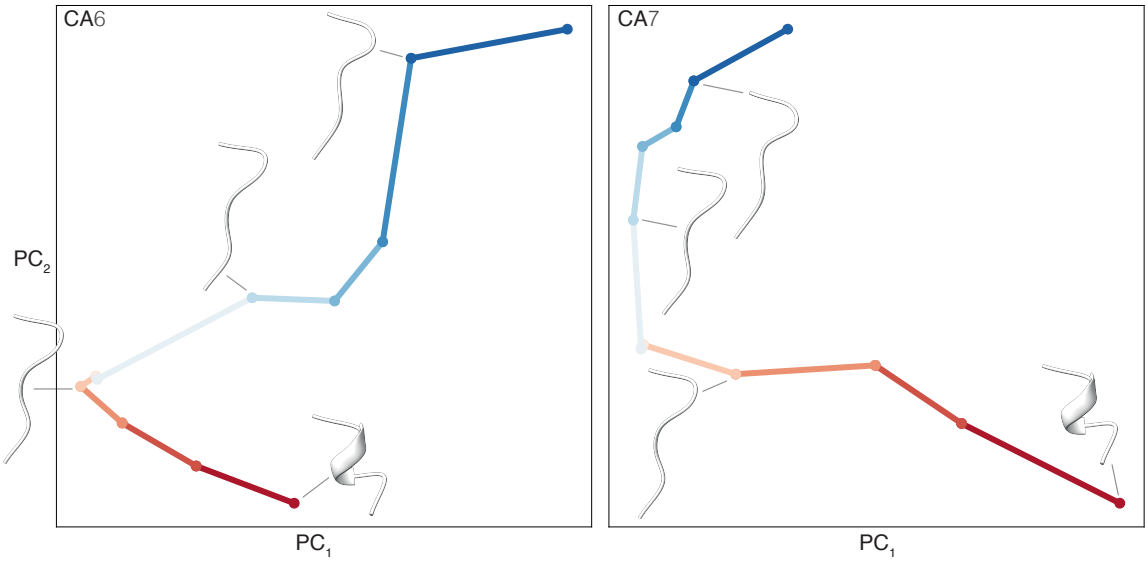

Figure A.8: **Structure of the MACE embedding space for the unfolding helix from Fig. A.12**. Depicted are Mace embeddings of the backbone CA atoms of residues 6 and 7 projected onto the first two principal components. Observe an essentially one-dimensional structure of the trajectory in the PCA space.

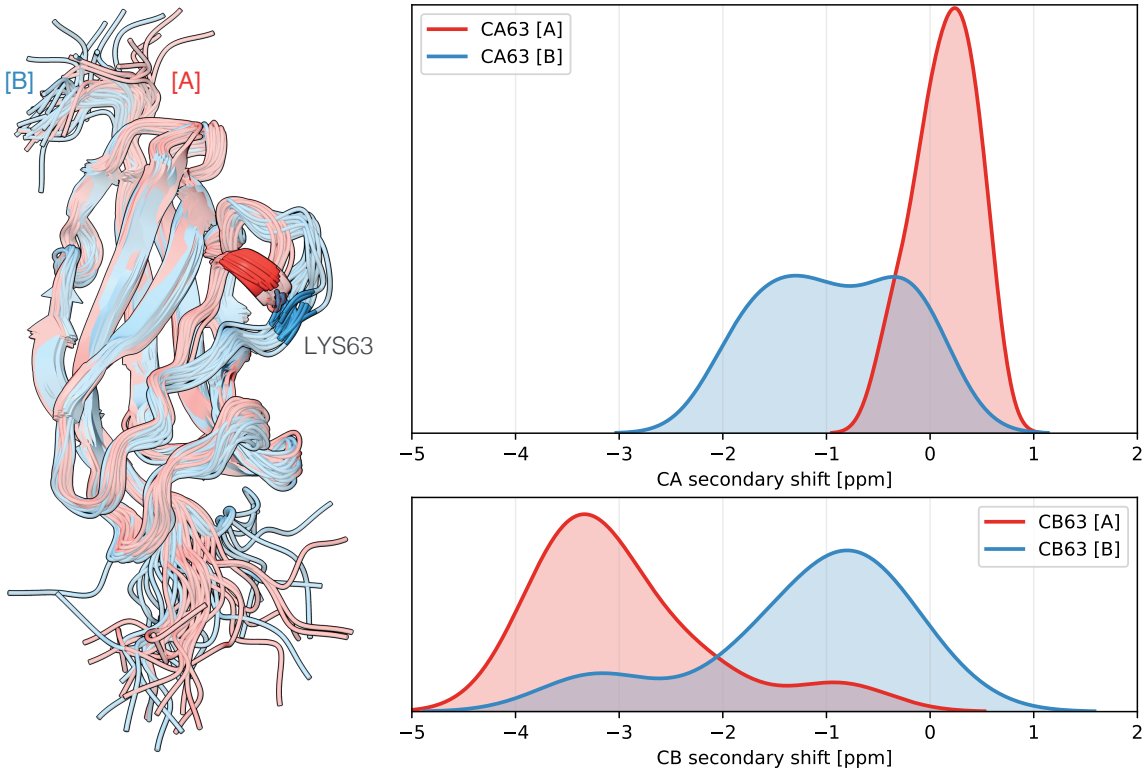

Figure A.9: **Distribution of the secondary chemical shifts of the two stable conformers of `4OLE`.** Left: 100 ns MD simulations (Rosenberg et al., 2024) of the two conformers (marked as A (containing a helix) and B (containing a linearly structured loop) according to the original PDB annotation). Lysine 63 is highlighted. Right: secondary shift distributions of lysine 63 CA (top) and CB (bottom) atoms over the MD trajectory. Note that the helical conformer exhibits an expected excess in the CA secondary shift and a defect in the CB secondary shift.

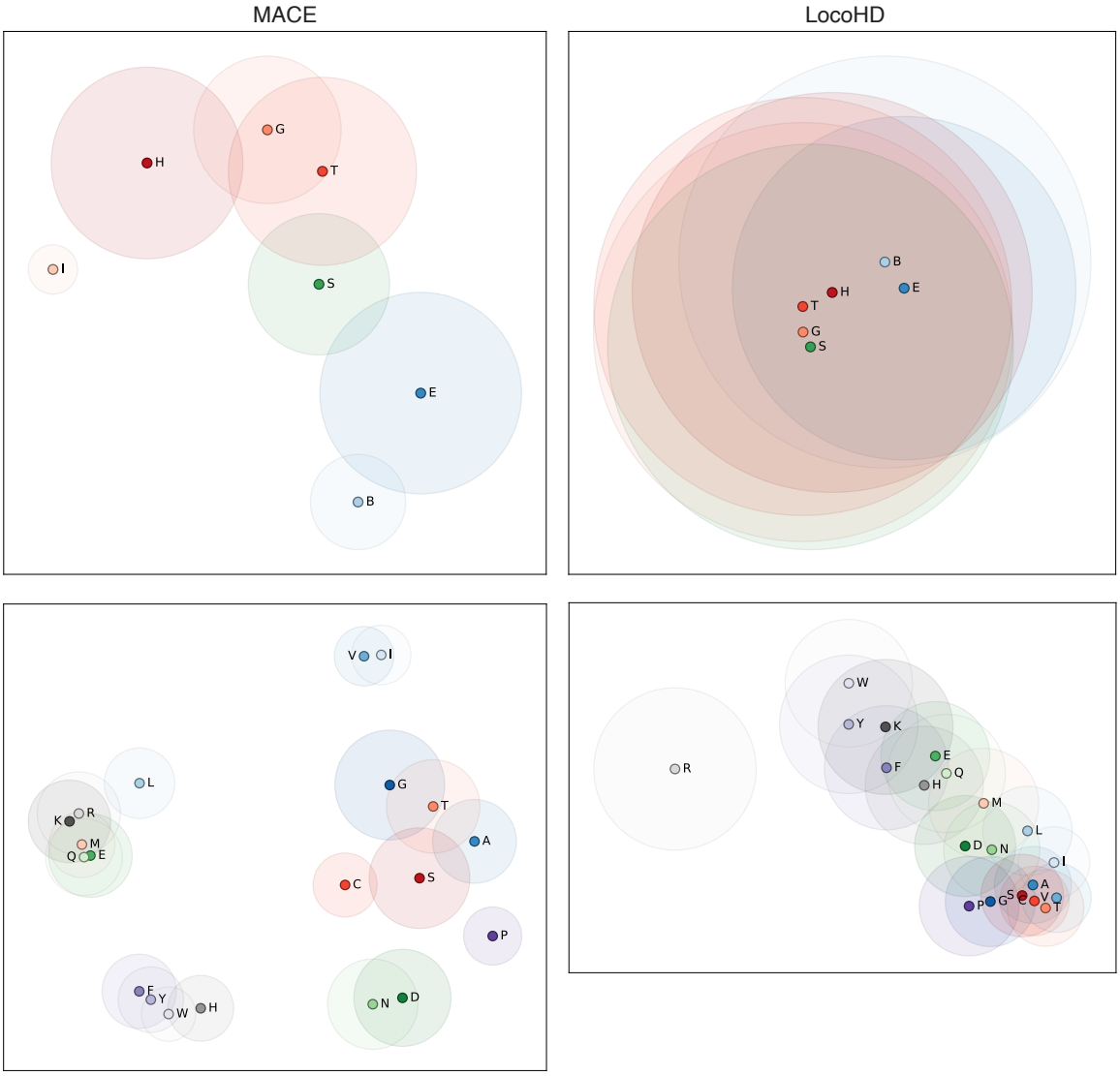

Figure A.10: **Similarity of local protein environments using MACE and LoCoHD.** In each plot, environments belonging to different class labels are represented as points while pairwise Euclidean distances approximate the dissimilarities as measured using the LoCoHD metric (Fazekas et al., 2024) and a metric between MACE likelihoods. Circles indicate each class variability as seen by the metric. Shown is conditioning by secondary structure (top row) and amino acid identity (bottom row). We conclude that MACE captures much better the similarities between related chemical and structural classes, even if LoCoHD has been explicitly designed for these tasks.

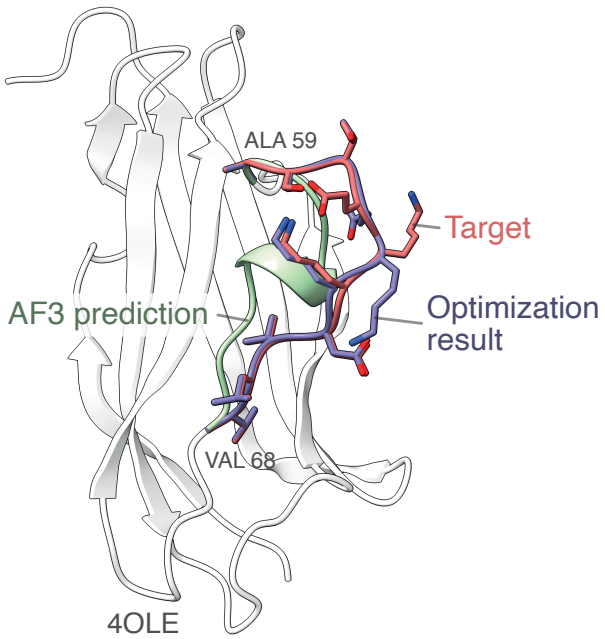

Figure A.11: **Optimized structure prediction using MACE descriptors.** Shown in green is the structure of the protein `4OLE:B` as predicted by AlphaFold3. The structure has an alternative conformation in the region $60 - 68$, (shown in red is the crystallographic structure from the PDB), whose MACE environment descriptors were used to optimize the AlphaFold prediction. The solution of the optimization problem is depicted in pink.

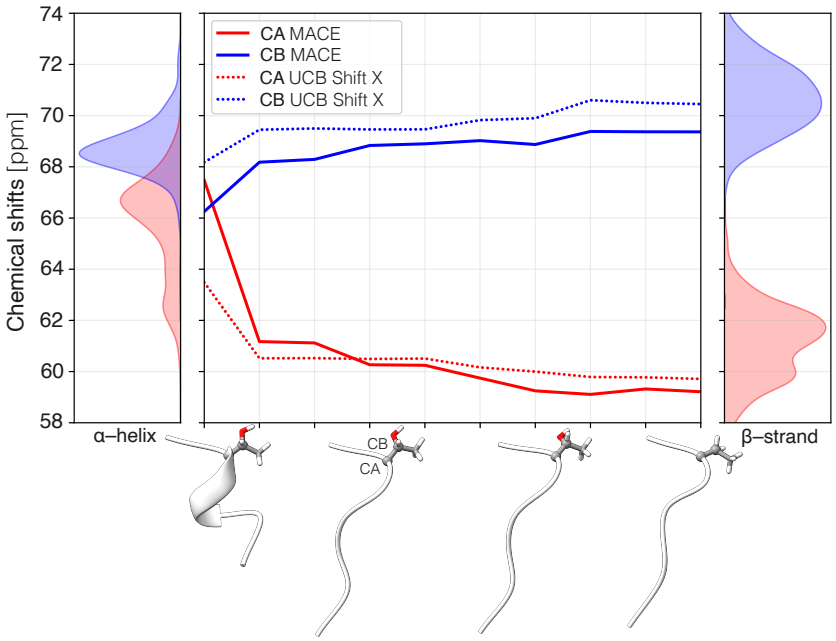

Figure A.12: **Simulated molecular dynamics trajectory of a helix unfolding into a strand.** Depicted are the chemical shifts of the CA and CB atoms in the middle of the helix as predicted by Mace and UCBShift2-X. Marginal plots show the experimentally-determined chemical shift distributions of the same atoms in helices (left) and strands (right). The known tendency of CA's to have smaller shifts in strands than in helices and the opposite CB tendency is clearly visible.

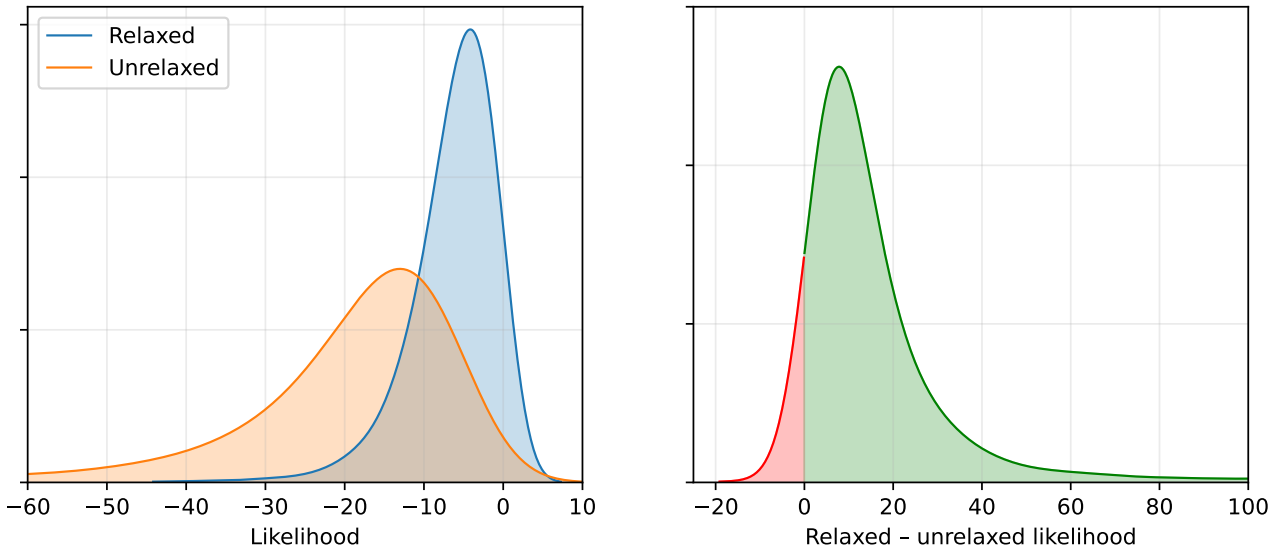

Figure A.13: **MACE-based likelihood captures structurally subtle distribution shifts in protein environments.** Left panel: the histogram of likelihoods of CA MACE representations of 26000 protein environments curated from 225 protein structures in the test set, before (red) and after (blue) relaxing the protein structures with an Amber99 force-field (Hornak et al., 2006). Right panel: the relative likelihood of the same environment before and after relaxation.

B. Quantitative Results & Tables

Table B.1: **Secondary structure prediction accuracy.** Reported are precision, recall, and F1 scores, each calculated with respect to the ground truth DSSP secondary structure classes. The best-performing model for each metric is indicated in **bold**.

			AFM-Features + GNN (Ours)			
Metric	**Class**	**LocoHD**	**Egret**	**MACE**	**OrbNet**	**AIMNet**
Precision [%]	Weighted Mean	76.947	95.685	**95.694**	92.500	94.716
	Mean	76.424	**95.540**	95.528	92.233	94.522
	α-helix	80.493	96.466	**96.657**	93.900	95.762
	β-strand	75.623	**96.248**	95.969	93.749	95.494
	Other	73.158	93.905	**93.958**	89.049	92.309
Recall [%]	Weighted Mean	77.334	**98.636**	95.717	92.574	94.752
	Mean	75.532	**98.508**	95.355	91.977	94.335
	α-helix	87.227	**97.546**	97.522	95.548	96.805
	β-strand	89.035	**97.967**	97.947	96.276	97.481
	Other	50.333	90.512	**90.597**	84.106	88.720
F1 [%]	Weighted Mean	76.323	**98.633**	95.693	92.509	94.720
	Mean	75.048	**98.552**	95.427	92.073	94.412
	α-helix	83.725	97.003	**97.087**	94.717	96.281
	β-strand	81.783	**97.100**	96.948	94.996	96.477
	Other	59.636	92.177	**92.247**	86.507	90.479

Table B.2: **Amino acid prediction accuracy.** Reported are precision, recall, and F1 scores, each calculated with respect to the ground truth amino acid class. The best-performing model for each metric is indicated in **bold**. Unlike secondary structure prediction, per-class performance is not reported here due to the large number (20) of amino acid types.

			AFM-Features + GNN (Ours)			
Metric	**Class**	**LocoHD**	**Egret**	**MACE**	**OrbNet**	**AIMNet**
Precision [%]	Weighted Mean	97.794	**99.129**	98.448	95.967	98.665
	Mean	98.012	**98.758**	98.031	95.090	98.407
Recall [%]	Weighted Mean	97.805	**99.133**	98.431	95.955	98.664
	Mean	97.699	**98.932**	98.262	95.072	98.457
F1 [%]	Weighted Mean	97.785	**99.130**	98.435	95.959	98.663
	Mean	97.839	**98.842**	98.137	95.078	98.431

Table B.3: **Acid Dissociation constant (pK_a) prediction accuracy.** Reported are mean absolute errors (MAE) relative to the pK_a values estimated using the Poisson-Boltzmann solver PypKa (Reis et al., 2020) on four amino acids. The best-performing model for each metric is indicated in **bold**.

				AFM-Features + GNN (Ours)		
Residues	**# Samples**	**PropKa**	**pKa-ANI**	**MACE**	**OrbNet**	**AIMNet**
Glutamic acid	12592	0.551	0.445	0.306	0.306	**0.265**
Aspartic acid	9841	0.469	0.473	0.280	0.284	**0.267**
Lysine	12009	0.393	0.401	0.320	0.282	**0.270**
Histidine	1182	0.561	0.426	0.424	0.441	**0.380**

Table B.4: **Chemical shift prediction accuracy**. Reported is the mean absolute error (MAE), in ppm, relative to the experimentally measured chemical shift values from RefDB (Zhang et al., 2003) for different atoms. The best-performing model for each metric is indicated in **bold**.

				AFM-features + GNN (Ours)			
Atom	**UCB Shift**	**LocoHD**	**ESM**	**MACE**	**Egret**	**Orb**	**AIMNet**
HA	**0.165**	N/A	0.247	0.180	0.176	0.200	0.201
H	0.324	N/A	0.402	**0.300**	0.295	0.333	0.325
CA	0.653	1.333	0.872	**0.584**	0.599	0.683	0.660
N	1.891	2.877	2.455	**1.642**	1.667	2.015	1.875
CB	0.758	1.140	0.923	0.716	**0.705**	0.792	0.780
C	0.827	1.297	1.011	0.744	**0.743**	0.847	0.824
CG	0.595	0.647	0.690	0.567	**0.563**	0.638	0.611
CD	0.615	0.512	0.550	0.456	**0.448**	0.521	0.482
CD2	1.116	1.195	1.335	1.035	1.085	1.201	1.749
CG2	0.773	0.833	0.927	**0.713**	0.720	0.760	0.742
CE	0.519	2.295	0.568	0.447	**0.436**	0.451	0.446

Table B.5: **Atomistic Foundation Model (AFM) metadata**. Listed are the model type, dimensionality, geometric properties of the representations, and training dataset – these include QM9 (Ramakrishnan et al., 2014), MD22 (Chmiela et al., 2023), ANI-1x (Smith et al., 2018), GDB13 (Cheng et al., 2019), DrugBank (Law et al., 2014), QM24 (Ruddigkeit et al., 2012) – for AFMs in Section 2.

Model Name	**Dimension**	**Properties**	**Dataset**
MACE large	448	S_n & $O(3)$ invariant	QM9 , MD22, ANI-1x
OrbNet-v2	256	S_n & $E(3)$ invariant	QM9, GDB13, DrugBank
AIMNet2	256	S_n & $E(3)$ invariant	ANI-1x
Egret1	384	S_n & $O(3)$ invariant	QM9, MD22 , ANI-1x, QM24

Table B.6: **Protein Secondary Structure Codes.** Summary of the standard single-letter codes used to represent protein secondary structure elements, as defined by the DSSP classification (Frishman & Argos, 1995).

Symbol	**Name**
H	α-helix
T	Turn
G	3_{10}-helix
I	π-helix
E	β-strand (extended)
B	Isolated β-hridge
S	Bend
–	Coil

Table B.7: **Secondary Structure Prediction Ablation.** Evaluation of secondary structure prediction accuracy when using a single Egret atom descriptor compared to using multiple Egret atom descriptors. The best-performing model for each metric is indicated in **bold**.

Metric	**Class**	$\mathcal{A}_a = \{\mathrm{CA}\}$	$\mathcal{A}_a = \{\mathrm{C, CA, CB, N, H, HA}\}$
Precision [%]	Weighted mean	94.822	**95.685**
	Mean	94.679	**95.554**
	α-helix	95.809	**96.466**
	β-strand	95.232	**96.248**
	Other	92.997	**93.905**
Recall [%]	Weighted mean	94.853	**98.636**
	Mean	94.552	**98.508**
	α-helix	96.759	**97.546**
	β-strand	97.971	**99.467**
	Other	88.925	**90.512**
F1 [%]	Weighted mean	94.817	**98.633**
	Mean	94.593	**98.552**
	α-helix	96.282	**97.003**
	β-strand	96.582	**97.103**
	Other	90.915	**92.177**

Table B.8: **Amino acid prediction ablation.** Evaluation of amino acid prediction accuracy when using a single Egret atom descriptor compared to using multiple Egret atom descriptors. The best-performing model for each metric is indicated in **bold**.

Metric	**Class**	$\mathcal{A}_a = \{\mathrm{CA}\}$	$\mathcal{A}_a = \{\mathrm{C, CA, CB, N, H, HA}\}$
Precision [%]	Weighted mean	95.897	**99.129**
	Mean	94.327	**98.758**
Recall [%]	Weighted mean	95.912	**99.133**
	Mean	94.069	**98.932**
F1 [%]	Weighted mean	95.629	**99.135**
	Mean	93.765	**98.842**

C. Extended Background

Chemical shifts. The resonance frequency of a nuclear spin, ν, depends directly on the splitting of its energy levels in a magnetic field. This is given as $\nu \propto \gamma B$, where B is the strength of the magnetic field that the nucleus experiences and γ is the gyromagnetic ratio of the nucleus, a characteristic of an atom's nucleus (e.g., ^{1}H, ^{13}C). In Nuclear Magnetic Resonance (NMR) experiments, B is primarily determined by the externally applied magnetic field B_0, but at the level of an observed atom it is slightly perturbed by local magnetic fields ΔB induced by the surrounding electron cloud. These perturbations, which are typically several orders of magnitude weaker than B_0, cause small shifts in the resonance frequency—known as chemical shifts. The chemical shift reflects subtle variations in the local electronic environment and is therefore highly informative about molecular structure and composition. The chemical change of an atom is often reported in relative terms as $\delta = 10^6 \cdot \frac{\nu - \nu_0}{\nu_0}$, where ν is the observed resonance frequency and ν_0 is the resonance frequency of the same type of nucleus in a reference compound – commonly the ^{1}H signal of 2,2-dimethyl-2-silapentane-5-sulfonate (DSS) is used for the ^{1}H chemical-shift scale. Although dimensionless, chemical shifts δ are conventionally expressed in parts per million (ppm), hence the 10^6 scaling factor in the definition.

The electronic environment that gives rise to chemical shifts is influenced by factors such as the nature of chemical bonds, the identities of neighboring atoms, and their three-dimensional spatial arrangement. Additionally, dihedral angles[2] play a significant role in modulating electron density around nuclei and thereby affect chemical shifts. Other factors that influence the chemical shift include hydrogen bonding interactions, ligand binding, or proximity to solvent molecules. Therefore, chemical shifts provide site-specific information about the local environment at the atomic-level. The chemical shift reports the time-averaged environment over all motions on time scales shorter than milliseconds. Chemical shifts are particularly valuable for inferring atomic connectivity, identifying functional groups, and detecting conformational changes, due to their sensitivity to fine variations in electronic structure (Claridge, 2016; Günther, 2013). Accurate chemical shift prediction plays a critical role in applications like automated NMR resonance assignment, molecular structure determination and validation, and the analysis of complex chemical systems (Wishart, 2011; Bermel et al., 2015). Despite this, reliable chemical shift prediction remains a challenge due to the sensitivity of chemical shifts to nuanced changes in molecular environment (Kuprov et al., 2007; Case, 2013).

ESM-based descriptors (Lin et al., 2023). Evolutionary Scale Modeling (ESM) is a family of transformer-based protein language models trained on a large-scale dataset of protein amino acid sequences using masked language modeling, similar to BERT (Devlin et al., 2019) for natural language. Unlike traditional methods that rely on multiple sequence alignments (MSAs) or co-evolutionary profiles, ESM models learn context-aware representations directly from amino acid sequences, enabling them to capture biochemical, structural, and evolutionary patterns within a protein sequence without explicit alignment.

Given an input amino acid sequence $\mathbf{a}$ of length L, a pre-trained ESM model produces a d-dimensional representation for each amino acid. These embeddings capture the identity of each amino acid and its contextual dependencies within the sequence. In this work, we use per-residue representations from the ESM-2 model (output from the 33rd layer of the 2B parameter model) as local environment descriptors, leveraging the model's ability to capture the evolutionary and biochemical context for each residue.

MACE and Egret descriptors (Batatia et al., 2022; Wagen et al., 2025). Message passing neural networks (Gilmer et al., 2017; Schütt et al., 2017) typically exchange messages between pairs of nodes in a graph. In contrast, the MACE family of models (MACE, Egret) generalizes this framework by learning equivariant messages involving a system of n (order parameter) nodes (or atoms) at a time. This enables richer geometric and chemical modeling by operating directly over higher-order interactions. Formally, for a given node at layer t of the graph, MACE first learns the following S_n-invariant and $O(3)$-equivariant features between pairs of atoms:

$$\mathbf{A}^{(t)}_{i,kl_3m_3} = \sum_{l_1m_1,l_2m_2} C^{l_3m_3}_{l_1m_1,l_2m_2} \sum_{j\in\mathcal{N}(i)} R^{(t)}_{kl_1l_2l_3}(r_{ji}) Y^{m_1}_{l_1}(\hat{\mathbf{r}}_{ij}) \sum_{\tilde{k}} W^{(t)}_{k\tilde{k}l_2} h^{(t)}_{j,\tilde{k}l_2m_2} \tag{2}$$

Here, R is a learnable radial function that encodes the distance between atoms i and j; Y is the spherical harmonic function that encodes the unit vector from atom i to atom j. The summation over the neighbors $\mathcal{N}(i)$ guarantees permutation invariance in the local aggregation. Additionally, $\mathbf{A}^{(t)}_{i,kl_3m_3}$ incorporates information from learned node embeddings at the previous layer using $\sum_{\tilde{k}} W^{(t)}_{k\tilde{k}l_2} h^{(t)}_{j,\tilde{k}l_2m_2}$. Lastly, $\sum_{l_1m_1,l_2m_2} C^{l_3m_3}_{l_1m_1,l_2m_2}$ (Clebsch-Gordan coefficients) ensure $O(3)$-

[2]The torsion angle about the bond between atoms B and C in a four-atom segment A–B–C–D

equivariance is preserved when radial and spherical features are combined. The indices $l_1, l_2, l_3, m_1, m_2, m_3$ generalize equivariant messages across higher-dimensional irreducible representations of $O(3)$. It must be noted that $\mathbf{A}^{(t)}_{i,kl_3m_3}$ captures pairwise directional and distance-based interactions, forming the building blocks for higher-order representations. Formally,

$$\mathbf{B}^{(t)}_{i,\eta_\nu kLM} = \sum_{\mathbf{lm}} C^{LM}_{\eta_\nu,\mathbf{lm}} \prod_{\xi=1}^{\nu} \sum_{\tilde{k}} w^{(t)}_{k\tilde{k}l_\xi} \mathbf{A}^{(t)}_{i,\tilde{k}l_\xi m_\xi}, \mathbf{lm} = (l_1m_1, \dots l_nm_n). \tag{3}$$

The above equation lifts the pairwise S_n (permutation-) invariant and $O(3)$-equivariant features in $\mathbf{A}^{(t)}_{i,kl_3m_3}$ to an n-body representation while preserving the geometric properties. The high-order tensor $\mathbf{B}^{(t)}_{i,\eta_\nu kLM}$ forms the core MACE representation. The update and readout steps are similar to regular message passing frameworks. As a special case, we set $L = 0$ (scalar representation of $O(3)$) in Equation 3, we project the equivariant features to a invariant representation of $O(3)$ via scalar coupling. Hence,

$$\mathbf{B}^{(t)}_{i,\eta_\nu k,00} = \sum_{\mathbf{lm}} C^{00}_{\eta_\nu,\mathbf{lm}} \prod_{\xi=1}^{\nu} \sum_{\tilde{k}} w^{(t)}_{k\tilde{k}l_\xi} \mathbf{A}^{(t)}_{i,\tilde{k}l_\xi m_\xi}, \mathbf{lm} = (l_1m_1, \dots l_nm_n). \tag{4}$$

Here $C^{00}_{\eta_n,\mathbf{lm}}$ would enforce scalar coupling of lower-order equivariant components. In this work, we will use descriptors (448 dimensional) that are invariant to both S_n and $O(3)$ transformations by setting $L = 0$.

AIMNet2-based descriptors (Zubatyuk et al., 2019). AIMNet constructs Atomic Feature Vectors (AFVs) by iteratively embedding the local atomic environments and updating atomic representations using geometric information. Given the atomic coordinates $\mathbf{R}$ and atomic number Z, an initial atomic feature vector $\mathbf{A}^{(0)}_i$ is derived for atom i. The initial embeddings are parametrized as learnable vectors that depend on atomic number Z_i of each atom.

The local environment of atom i is encoded using symmetry functions that capture the spatial arrangement and types of neighboring atoms. This produces an atomic environment vector (AEV).

$$\mathbf{G}^{(t)}_i = \text{AEV}(\{\mathbf{R}_i - \mathbf{R}_j, Z_j\}_{j\in\mathcal{N}(i)}). \tag{5}$$

An AEV is constructed to be invariant to permutations (S_n) and 3D translations and rotations ($E(3)$). This is typically via a message-passing framework. Next, the geometry and atom-type information are fused by computing an interaction descriptor

$$\mathbf{f}^{(t)}_i = \text{MLP}\left(\sum_{j\in\mathcal{N}_i} \mathbf{G}^{(t)}_{ij} \otimes \mathbf{A}^{(t)}_j\right) \tag{6}$$

This vector encodes the influence of neighboring atoms on atom i in a way that is independent of the number of chemical species. Lastly, the atomistic feature vectors $\mathbf{A}^{(t)}_j$ and $\mathbf{f}_i$ are combined to form an updated atomistic feature vector

$$\mathbf{A}^{(t+1)}_i = U(\mathbf{A}^{(t)}_i, \mathbf{f}^{(t)}_i). \tag{7}$$

This process is repeated over multiple iterations, enabling information to propagate beyond the local neighborhood and capture non-local effects such as charge redistribution and polarization. After several refinement steps, AIMNet produces AFVs of fixed dimensionality (256), which are used to predict various molecular and atomic properties. Training is performed in a multitask setting, where the network jointly learns to predict a variety of molecular and atomic properties—including energies, forces, partial charges, and dipole moments—using shared representations to improve generalization. The model is trained on large datasets of quantum mechanical calculations, with regularization techniques to ensure stability and prevent overfitting.

OrbNet-based descriptors (Qiao et al., 2020). In similar spirit to semi-empirical molecular force-fields, OrbNet constructs feature vectors by leveraging symmetry-adapted atomic orbitals (AO), which capture local electronic structure information obtained from quantum calculations while respecting the symmetries inherent in molecular systems. Using mean-field quantum mechanics, OrbNet derives overlap matrices $\mathbf{S}$, Fock matrix $\mathbf{F}$, density matrix $\mathbf{D}$, and the Hamiltonian $\mathbf{H}$. Each matrix element $O^{\nu\mu}$ corresponds to interactions between atomic orbitals ν and μ, and blocks $\mathbf{O}_{ij}$ represent interactions between atoms i and j.

These matrices produce features that are S_n invariant and $E(3)$ invariant by projecting these orbital block matrices into a symmetry-adapted basis. The atom-wise feature vector is then constructed by aggregating contributions from its neighboring orbital interactions

$$\mathbf{f}_i = \sum_j \text{MLP}(\mathbf{O}_{i,j}). \tag{8}$$

In OrbNet, the atomic structure is represented as a graph where nodes correspond to an atom (initialized using $\mathbf{f}_i$) and edges model interatomic relationships. The model employs the following message passing scheme

$$\mathbf{h}_i^{(t+1)} = U^{(t)}\left(\mathbf{h}_i^{(t)}, \sum_{j\in\mathcal{N}(i)} M^{(t)}(\mathbf{h}_i^{(t)}, \mathbf{h}_j^{(t)}, \mathbf{O}_{ij})\right), \tag{9}$$

where $\mathbf{h}_i^{(t)}$ is hidden feature of atom i at layer t, $\mathbf{O}_{ij}$ encodes the orbital-based edge features derived from orbital feature matrices, $M^{(t)}$ is the message function, and U is the update function for node embeddings. After multiple layers of message passing, the final atomic features $\mathbf{f}_i \in \mathbb{R}^{256}$ serves as a descriptor of an atom's local environment. This formulation enables OrbNet to encode both local chemical environments and long-range quantum effects efficiently. In contrast to models like MACE, which explicitly encode geometric equivariance and higher-order correlations, OrbNet uses pairwise orbital-based features and rely on message passing to capture more complex interactions.

LoCoHD (Fazekas et al., 2024) is a method for quantifying chemical and structural differences between protein environments. Unlike alignment-based or purely geometric measures, LoCoHD characterizes each local environment as *a distribution of chemical "primitive types"* such as atom types or residue centroids within a specified radius. The similarity between two environments is then measured using the Hellinger distance between their respective distributions. In this paper, to test the effectiveness of LoCoHD descriptors in representing local environments, we construct *LocoHD embeddings* using its representation of a local environment while computing the similarity metric. Each embedding is computed by aggregating statistics of the primitive types, weighted by their distance from the central residue. This results in a purely structural and chemical descriptor of the local environment.

D. Data curation

The dataset used within the scope of this paper was sourced from RefDB (Zhang et al., 2003) which is a subset of the Biomolecular Magnetic Resonance Databank (BMRB) (Hoch et al., 2023). RefDB is a curated list of BMRB entries with calibrated chemical shifts. We select monomers from the RefDB and perform sequence redundancy filtering with `mmseqs` via,

```
mmseqs easy-cluster sequences.fasta -min-seq-id 0.5 -c 0.8 -cov-mode 5,
```

which results in sequence clusters. We then split the clusters to define train and test sets. Our final dataset consists of 1048 BMRB entries in RefDB (225 test + 823 train), from where we collected the amino acid sequences, and the experimental chemical shifts. The structures are obtained by running AlphaFold2 (Jumper et al., 2021) via OpenFold (Ahdritz et al., 2024). Using AlphaFold offers two advantages: (i) it avoids structural inconsistencies that may be inherent in crystal structures, and (ii) it allows predicting shifts for structures that do not have a crystal structure deposited to the PDB. The local protein environments are obtained from each protein structure are extracted as described in Algorithm 1. We obtain the random-coil chemical shifts used in our shift predictor from UCBShift2 (Ptaszek et al., 2024).

E. Environment construction and AFM embedding

This section elaborates on the amino-acid centric environment construction algorithm discussed in Section 3.

Algorithm 1 Amino-Acid Centric Environment Extraction

Input: Protein dataset $\mathcal{D}$, radius $r_{\text{max}} = 5\text{Å}$
$\mathcal{E} \leftarrow \emptyset$
for each structure $\mathcal{X} \in \mathcal{D}$ **do**
  for each amino acid $a \in \mathcal{X}$ **do**
    for each atom $\mathbf{x}_i \in a$ **do**
      $\mathcal{X}_a \leftarrow \{a\}$ $\qquad$ $\mathcal{X} \in \mathbb{R}^{N_{\text{atoms}} \times 3}, \mathbf{x}_i \in \mathbb{R}^3$
      for each amino acid $a' \in \mathcal{X}$ **do**
        if $\exists \mathbf{x}_j \in a'$ s.t. $\|\mathbf{x}_i - \mathbf{x}_j\| \leq r_{\text{max}}$ **then**
          $\mathcal{X}_a \leftarrow \mathcal{X}_a \cup \{a'\}$ $\qquad$ All atoms in a'
        end if
      end for
      $\mathcal{E} \leftarrow \mathcal{E} \cup \mathcal{X}_a$
    end for
  end for
end for
Return $\mathcal{E}$

F. Predicting amino acid, secondary structure, and pK$_\text{a}$

In what follows, we provide additional details for experiments and results described in Section 4.

F.1. Secondary Structure and Amino Acid Prediction

Target variables. In these tasks, the target space of the target y is categorical variable.

- In **amino acid prediction**, the target variable y holds one of the following categories: $\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$, each representing a standard twenty amino acids (cf. (Kawashima & Kanehisa, 2000) for further explanation).

- In **secondary structure prediction**, the target variable $\mathbf{y}$ can assume one of the following categories: $\{H, E, O\}$, where H denotes an α-helix, E denotes a β-strand, and O corresponds to all other secondary structure types (see Table B.6). Although more secondary structure types exist, we are primarily interested in α-helices and β-strands due to their structural prominence.

Loss function. To train models for these tasks, we employ a *cross entropy loss* written as

$$\mathcal{L}(\hat{\mathrm{y}}_i, \mathrm{y}_i) = -\sum_{c=1}^{|\mathrm{y}|} y_{ic} \log \left(\frac{\exp(\hat{y}_{ic})}{\sum_{j=1}^{|\mathbf{y}|} \exp(\hat{y}_{ij})} \right), \tag{10}$$

where $\hat{\mathrm{y}}_i$ is the *predicted logit* for each class and y_i is the *one-hot encoding* of the true class label for sample i.

Evaluation metrics. To evaluate the model's performance, we use the following metrics.

- *Precision*: For a given class i, precision is defined as

$$\text{Precision}_i = \frac{\text{True Positive}_i}{\text{True Positive}_i + \text{False Positive}_i}$$

- *Recall*: For a given class i, recall is defined as

$$\text{Recall}_i = \frac{\text{True Positive}_i}{\text{True Positive}_i + \text{False Negative}_i}$$

- *F1 score*: For a given class i, the F1 score is the harmonic mean of its Precision and Recall:

$$\text{F1}_i = 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

For secondary structure prediction we report the precision, recall, and F1 score for each class in y. In addition, we report an unweighted average and a weighted average (mean weighted by number of instances in each class) for the respective tasks.

Ablation study. We conduct an ablation study on the set of atoms $\mathcal{A}_a$ to evaluate the effect of including descriptors from multiple atoms within amino acid a on secondary structure and amino acid prediction performance. Specifically, two configurations are compared: one where $\mathcal{A}_a = \{\text{C, CA, CB, N, H, HA}\}$, and a reduced version $\mathcal{A}_a = \{\text{CA}\}$. For this ablation, identical models are trained using Egret descriptors, varying only the atom set. Egret descriptors are used because as shown in Tables B.1 and B.2, they consistently outperform other atom-level descriptor types. The ablation results for secondary structure prediction are presented in Table B.7 and for amino acid prediction in Table B.8. Clearly, incorporating a larger set of atoms in $\mathcal{A}_a$ leads to a substantial improvement in prediction accuracy.

F.2. Acid dissociation constant (pK_a) prediction

Loss function. We treat pK_a prediction as a regression problem. To train pKa regressors, we train a graph neural network to minimize the mean absolute error (MAE) as follows,

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|, \tag{11}$$

where N is the number of samples in the dataset, y_i^* is the *groundtruth* pK_a value of a protonated site in a protein obtained by solving the Poisson-Boltzmann equation from (Reis et al., 2020), and $\hat{y}_i$ is the *predicted* pK_a.

Evaluation metrics. To quantitatively evaluate the accuracy of the model, we report the mean absolute error measured on the test set.

F.3. Model Architecture

Across all three tasks, we used a consistent model architecture with little changes in hyperparameters. To incorporate the spatial and relation information between atomistic features, we constructed *a fully connected graph between the atoms of the same residue*. We propose to use a graph convolution network (GCN) (Kipf & Welling, 2016) to predict the specific class type. The update rule at layer $l \in \mathbb{N}$ is defined as,

$$\mathbf{X}^{(l+1)} = \sigma(\bar{\mathbf{A}}\mathbf{X}^{(l)}\mathbf{W}^{(l)}) \tag{12}$$

Here, $\bar{\mathbf{A}} = (\mathbf{I} + \mathbf{A})$ is the normalized adjacency matrix that serves as a low-pass filter when aggregating information over the neighbors. $\mathbf{A}$ is the graph's adjacency matrix, and $\mathbf{I}$ is an identity matrix. Also, $\mathbf{W}^{(l)}$ is the learnable weight matrix to transform the features at layer l, $\mathbf{X}$ is the node feature matrix defined for layer l of the GNN. We used SiLU (Sigmoid Linear Unit) (Elfwing et al., 2018) as the non-linear activation function and we applied layer normalization after each graph convolution operation to stabilize training. The operation in Eq. 12 is repeated for $L = 10$ layers. The initial input $\mathbf{X}^{(0)}$ consists of the atomistic features for each atom. At the final layer ($l = L - 1$), in the case of classification, the network outputs a vector of shape $|\mathrm{y}|$ representing class-wise logits, and the network outputs a scalar for regression tasks. Lastly, we use dropout (Srivastava et al., 2014) for every intermediate layer to prevent overfitting. Below is the table of hyper-parameters used for these experiments.

- $\mathcal{A}_a = \{\text{C}, \text{CA}, \text{CB}, \text{N}, \text{H}, \text{HA}\}$

- Batch size: 256
- Number of Epochs: 800
- Optimizer
  - Type: Adam (Kingma & Ba, 2014)
  - Learning Rate: 1×10^{-4}
- Learning Rate Scheduler
  - Type: Step Learning Rate
  - Decay Factor (γ): 0.5
  - Step Size: 10
- Number of GCN layers: 10
- Dropout probability: 0.3
- Hidden dimension of GCN (per layer): 128

In addition, we employed gradient clipping with a maximum norm of 5, an exponential moving average (EMA) weighted model with a decay rate of 0.999, and a StepLR learning rate scheduler incorporating 7500 warm-up steps and a step size of 10.

F.4. Computation details

All experiments were performed on a single NVIDIA H100 GPU. Individual runs required one hour to complete, with a maximum GPU memory usage of approximately 15 GB.

G. Using likelihoods to define environment similarity measure

We employ the conditional likelihoods defined in Section 5 to define a similarity metric to compare protein environments.

G.1. Similarity metric

We analyzed a large collection of protein chemical environments by comparing them using MACE embeddings. For each attribute—such as amino acid type or secondary structure—we estimated a likelihood around each embedding via kernel density estimation with a fixed bandwidth as described in Section 5. These likelihoods were converted into distances using a scaled negative exponential, where higher similarity corresponded to smaller distances. Since each pairwise distance arises from a distribution—based on multiple environments belonging to the same attribute—we applied stochastic multidimensional scaling (MDS) (Rosenberg et al., 2022; Boyarski & Bronstein, 2021), which takes into account both the mean and standard deviation of these distances. The resulting embeddings revealed meaningful clustering that not only separated different amino acid types and secondary structures but also captured finer relationships within them. Chemically or structurally related amino acids and secondary structures exhibited higher similarity, demonstrating that the MACE-based comparison effectively reflects the underlying chemical and structural properties. Additional details and visualizations are provided in Figure A.10.

H. Details of the chemical shift prediction experiments

Target variable. The goal of the chemical shift prediction is to predict the calibrated chemical shift y of an atom of interest given the environment. We train our models to predict the *secondary shift*, i.e. the difference in the experimental and random-coil chemical shifts, given the sequence.

Loss function. We train our shift prediction models to minimize the mean absolute shift prediction error, defined as

$$\mathcal{L}_{\mathrm{L1}} = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \,, \tag{13}$$

where N is the number of data points, y_i and $\hat{y}_i$ are the groundtruth and predicted chemical shifts, respectively.

Evaluation metrics. To quantitatively evaluate the accuracy of the model, we report the mean absolute error measured on the test set.

H.1. Model architecture

We used the same architecture as before (Appendix F.3) with slightly different hyperparameters as described below.

- $\mathcal{A}_a = \{\mathbf{b}\}$, where $\mathbf{b}$ is the atom type for which we are predicting the chemical shift value.
- Batch size: 500
- Number of Epochs: 2000
- Optimizer
  - Type: Adam (Kingma & Ba, 2014)
  - Learning Rate: 1×10^{-4}
- Learning Rate Scheduler
  - Type: Step Learning Rate
  - Decay Factor (γ): 0.5
  - Step Size: 10
- Number of GCN layers: 5
- Dropout probability: 0.3
- Hidden dimension of GCN (per layer): 256

In addition, we employed gradient clipping with a maximum norm of 5, an exponential moving average (EMA) with a decay rate of 0.999, and a StepLR learning rate scheduler incorporating 7500 warm-up steps and a step size of 10.

H.2. Computation details

All experiments were performed on a single NVIDIA H100 GPU. Individual runs required three hours to complete, with a maximum GPU memory usage of approximately 15 GB.

I. Extended AFM interpretations

I.1. Inverting MACE embeddings

In a similar spirit to Maddipatla et al. (Maddipatla et al., 2025), we investigate whether MACE features can serve not only as representations of protein structures but also as tools to decode them. Specifically, we are interested in using the information contained in the descriptors to select between alternate conformers (altlocs) within a single protein. We therefore treat the MACE features as latent representations from which protein conformers can be reconstructed. As a test case, we examine the PDB structure `4OLE:B` resolved to 2.52 Å, which contains a region of 9 amino acids ($60 - 68$) modeled as a superposition of two alternate conformations – one helix and the other a loop. Starting from the helix, we iteratively optimize the positions to align the MACE features of the loop conformation, thereby assessing the potential of these descriptors for structure recovery and conformational transitions.

The optimization loss function quantifies the discrepancy between the descriptors of the current and target structures and is computed for each atom within the optimized region. In particular, we define each local atomic environment as all atoms

within a 5 Å radius around the atom of interest. The resulting descriptors capture both geometric and chemical context at multiple equivariant orders. The loss is composed of three terms, each corresponding to a different equivariant order:

$$\mathcal{L} = \alpha \left\| \mathbf{B}_{i,\eta_\nu k,00} - \mathbf{B}^{\text{target}}_{i,\eta_\nu k,00} \right\|_2 + \beta \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{B}^{i}_{i,\eta_\nu k,1M} - \mathbf{B}^{\text{target},i}_{i,\eta_\nu k,1M} \right\|_2 + \gamma \sum_{i} \left\| \mathbf{B}^{i}_{i,\eta_\nu k,2M} - \mathbf{B}^{\text{target},i}_{i,\eta_\nu k,2M} \right\|_F$$

where $\mathbf{B}^{i}_{i,\eta_\nu k,LM}$ and $\mathbf{B}^{\text{target}}_{i,\eta_\nu k,LM}$ denote the L-th order MACE embeddings of the optimized and target conformations, computed as described in Equation 3. For a given L, the Clebsch-Gordan coefficients $C^{LM}_{\eta_\nu,\mathbf{lm}}$ select the valid combinations of input angular momenta. Note that while the second and the third terms are not rotation-invariant, global alignment of the source configuration with the target ensures validity of the comparison.

This combined loss (weighted by α, β, γ) drives the structural optimization by progressively aligning the descriptor representations of the initial and target altloc conformations.

Fig. A.11 illustrates the result of this optimization. While the backbone atom positions are recovered with high accuracy, the side chains of glutamic acid (GLU) 62 and lysine (LYS) 63 adopt incorrect orientations. We hypothesize that this discrepancy arises because MACE features primarily capture the local chemical environment within a given radius. Within this range, distinct side-chain rotamers can produce similar local environments, resulting in ambiguous representations. As a result, the descriptors lack the resolution necessary to distinguish between different side-chain conformations.

These results suggest that while MACE features reliably capture the backbone geometry, are less sensitive to side-chain orientations, as distinct rotameric states can yield similar local environments. Recovering both backbone and side-chain configurations would require further exploration. Success in this direction could enable the use of MACE features (and other atomistic features) for coarse-grained protein modeling with minimal loss of structural fidelity.