Localized Reactivity on Proteins as Riemannian Manifolds: A Quantum-Inspired Geometric Model for Deterministic, Metal-Aware Reactive-Site Prediction

Haneul Park
Department of Biological Sciences, Seoul National University
Department of Mathematical Sciences, Seoul National University
phl0434@snu.ac.kr

Abstract

We introduce a unified framework for analysing molecular reactivity that centres on a geometric, quantum-inspired environment representation and a fully deterministic, metal-aware implementation. Proteins and protein complexes are treated as configurations in $\mathbb{R}^3 \times T$, and each residue p is mapped to an environment vector E_p that encodes its local geometric and physicochemical context. Motivated by Density Functional Theory, we construct a coarse-grained surrogate of local electron density around the Cα trace, augmented with metal- and phosphate-like fields, solvent exposure, curvature and neighbour geometry, and hypothesise that this local stencil retains enough information to separate reactive from non-reactive environments at residue resolution.

On top of this formalism we provide a deterministic Python pipeline that maps arbitrary PDB/mmCIF structures to fixed-dimensional environment vectors without any stochastic training or parameter fitting. To quantify how much functional signal this representation exposes, we treat the environment vectors as features and probe them with a simple, untuned Random Forest classifier under a site-grouped cross-validation protocol that prevents symmetry-related copies of the same residue or site from leaking across folds. Across representative case studies including GroEL/GroES, p53–DNA, Rubisco and SecA, the framework alone yields ROC–AUC values in the 0.8–0.9 range for many catalytic, metal-coordination, tight interface and hinge-support cores, while showing substantially weaker and noisier separation for labels tied to global fitness, diffuse allosteric networks or scaffolding roles.

Conceptually, this work should be viewed as defining an interpretable, physics-grounded representation rather than a task-specific predictor. Our contributions are: (i) a geometric and measure-theoretic formalisation of residue environments as an environment-vector manifold; (ii) a fully deterministic, metal-aware implementation that maps structural configurations into this space; and (iii) a unified evaluation protocol that uses a shallow classifier as a probe to characterise which classes of functional annotation become linearly or weakly non-linearly accessible in the environment space. Future work will extend this representation to time-resolved trajectories and multi-state interfaces and integrate it with evolutionary and dynamical information, as well as with topological data analysis on the environment manifold.

Keywords: protein manifold; environment vector; reactivity representation; differential geometry; quantum chemistry; deterministic framework; topological data analysis

Contents

1 Introduction **3**

2 Theoretical Framework **7**
2.1 Manifold Representation of Protein Structure . . . 7
2.2 Environment Vector E_p . . . 7
2.3 Fiber bundle for reactive states . . . 8

3 Methods **9**
3.1 Overview of the 1–9 Environment Pipeline . . . 9
3.1.1 Input Structures and Residue Representation (Step 1) . . . 9
3.1.2 Solvent Accessible Surface Area (SASA) (Step 1b) . . . 10
3.1.3 Residue Contact Graph and Fragment Definition (Step 2) . . . 10
3.1.4 Surrogate Promolecular Electron Density (B-mode) (Step 3) . . . 10
3.1.5 Optional Quantum-Corrected Density (A-mode) (Step 4) . . . 11
3.1.6 Density-Derived Local Descriptors and Element-Resolved Channels (Step 5) . . . 11
3.1.7 Charge-Derived Descriptors (Step 6) . . . 13
3.1.8 Local Integration Domains U_p (Step 7) . . . 14
3.1.9 Theoretical Environment Components (Step 8) . . . 15
3.1.10 Physicochemical 4D Residue Embedding (Step 8, component 6) . . . 16
3.1.11 Geometric Neighbor Augmentation (Step 9 extra features) . . . 17
3.1.12 Final Environment Vectors (Step 9) . . . 18
3.1.13 Outputs and Reproducibility . . . 19
3.2 Supervised evaluation of residue-level feature representations . . . 19
3.2.1 Residue-level feature matrix . . . 19
3.2.2 Definition of residue-level labels . . . 19
3.2.3 Random forest classifier and cross-validation protocols . . . 21
3.2.4 Permutation-based shuffle test for statistical significance . . . 24

4 Results **25**
4.1 Case study: Rubisco large/small subunits (4RUB) . . . 25
4.2 Case study: GroEL/GroES chaperonin complex (2C7C) . . . 29
4.3 Case study: p53–DNA complex (1TSR) . . . 33
4.4 case study: SecA ATPase motor and preprotein clamp (1TF5) . . . 36

5 Discussion **39**
5.1 Where the representation is strong . . . 39
5.2 Where the representation is weak . . . 40
5.3 What the results mean from a representation-only perspective . . . 40
5.4 Practical implications for structural bioinformatics pipelines . . . 41

5.5 Limitations and future directions 41

A Formal Definition of the Artificial Manifold Class V' **42**
A.1 Overview 42
A.2 Sampling a Geometric Structure from V' 42
A.3 Environment Vectors in V' 43
A.4 Reactive Fiber Assignment in V' 46

B Mathematical foundations of the environment–vector space **47**

C Time-Coherent Reactivity **47**

1 Introduction

High-throughput structure determination and structural genomics have produced an abundance of experimentally resolved and predicted protein structures[1, 2], yet *functional* annotation still lags far behind[3]. Identifying catalytic, allosteric, or metal-coordination residues often requires extensive prior knowledge, labour-intensive mutagenesis, or expensive QM/MM simulations. In parallel, geometric and graph-based pipelines are increasingly able to highlight metal-binding residues, pockets, and interfaces from local descriptors[4, 5, 6, 7], but non-deterministic learning stages and complex software stacks can impede reproducibility and make it difficult to separate genuine physical signal from model-specific artefacts. There is thus a need for approaches that combine physical interpretability, deterministic behaviour, and computational efficiency while still exposing enough structure for downstream predictors to exploit.

In this work we take a representation-first perspective. We regard proteins, protein complexes, and (in principle) their conformational changes as objects embedded in a common ambient space $\mathbb{R}^3 \times T$, where $\mathbb{R}^3$ denotes physical space and T indexes time or external conditions such as ligand state. A single static structure can be viewed as a finite sample from a submanifold of $\mathbb{R}^3 \times T$, an MD trajectory traces out a curve or thin tube, and a set of interacting proteins corresponds to a collection of such submanifolds that intersect or approach each other in this space. Our goal is to equip this geometric picture with a basic, physically motivated notion of local "reactivity" that is defined uniformly for all residues and configurations, so that both intra-protein organisation and inter-protein contacts can be analysed within a single coordinate system. In this paper we instantiate the framework on amino-acid residues in proteins, but the construction is intended to extend to more general molecular systems at the level of atoms or functional groups.

Position relative to existing work

Structure-based local environments and functional sites. A long line of work has used local structural environments to annotate functional residues in proteins. Early methods such as FEATURE and related environment-vector approaches represent a residue-centred neighbourhood by hand-crafted physico-chemical descriptors and then transfer annotations by matching similar environments across structures [8]. Subsequent tools, including pocket-

and cleft-based pipelines such as ConCavity and related geometry–conservation hybrids, treat local surface patches as the basic unit and combine shape with sequence conservation to detect catalytic or binding sites [5, 9]. More recent deep-learning methods derive local representations from equivariant 3D networks or atomistic foundation models, using learned embeddings of residue-centred environments as powerful but opaque features for tasks such as chemical shift prediction or binding-site classification[6, 10].

Electron-density and conceptual DFT descriptors of reactivity. From the quantum chemistry side, conceptual Density Functional Theory (DFT) and electron-density analyses have established ground-state density and its derivatives as fundamental descriptors of molecular stability and reactivity [11]. In the enzymology literature, several studies have used charge density, Laplacians of the electron density, and Fukui-type indices at key atoms to rationalise catalytic activation and electrostatic preorganisation at specific active sites, typically using QM or QM/MM calculations in a small predefined region[12] These works provide physically grounded descriptors of local reactivity, but they are computationally expensive and are not designed to yield a deterministic, per-site representation over an entire protein or across many proteins.

Geometric and topological representations of protein structure. A third line of work treats proteins and molecular assemblies as geometric or topological objects. Topological data analysis (TDA) pipelines such as SINATRA Pro [13]and graph-based methods like PersGNN summarise ensembles of protein conformations with persistent homology [14] or related signatures, and then use these summaries in statistical or deep models to distinguish functional states or binding modes.

Recent studies interpret protein conformational space as a Riemannian manifold equipped with a metric derived from physical energy landscapes, enabling the analysis of global dynamics and nonlinear trajectories [15].

While these approaches successfully capture the macroscopic geometric evolution of proteins, translating these global dynamic descriptions into explicitly interpretable, density-inspired coordinates for the local reactivity of individual residues remains an open challenge.

Metal- and ligand-focused electron-density tools. Several specialised methods focus on metal ions and small-molecule ligands using experimental electron-density maps. For example, map-validation tools classify local density blobs as likely ligands or ions and flag inconsistencies between model and map[16]. More recent frameworks explicitly predict the identity and localisation of metal ions and waters in structural data[17, 18]. Such methods operate on local electron density and its neighbourhood, but they are tailored to specific species (ligands, metals, waters) rather than providing a unified representation of residue-level reactivity across catalytic, binding, and regulatory roles.

Position of the present work. The present framework sits at the intersection of these lines. Conceptually, we adopt the DFT-inspired view that electronic structure is encoded in electron density and external fields, but we work at the level of *surrogate* density fields

and coarse-grained descriptors that can be computed deterministically from standard structural data. Geometrically, we treat proteins and complexes as configurations embedded in $\mathbb{R}^3 \times T$ and use the resulting environment vectors to turn each configuration into a finite metric–measure space in a common "reactivity" coordinate system. Practically, we provide a single, fully deterministic pipeline that assigns to every residue in a structure a fixed-dimensional environment vector, without any stochastic training or task-specific tuning. Unlike task-optimised site predictors, our focus is not on squeezing out state-of-the-art performance for a particular endpoint, but on quantifying how much functional structure—catalytic, metal-coordination, DNA-contact, protein–protein interface, allosteric, and post-translational modification sites—is already accessible from this unified representation using only shallow classifiers. In the longer term, the same construction is intended to extend beyond proteins to more general molecular systems by changing the notion of "site" from residues to atoms or functional groups while keeping the underlying reactivity coordinate space.

Concretely, we represent the Cα trace as a one-dimensional curve embedded in $\mathbb{R}^3 \times T$ and endow a tubular neighbourhood with an induced surface metric. Around each residue p we define a local neighbourhood U_p and construct an *environment vector* $\mathrm{Ep}_p \in \mathbb{R}^k$ that aggregates coarse-grained information about local charge, surrogate density, curvature, solvent exposure, and residue identity, together with a small set of neighbour-based geometric moments and field terms. This design is inspired by the Hohenberg–Kohn principle, which states that the ground-state electron density determines electronic structure: instead of attempting to reconstruct the full density field, we view Ep_p as a low-dimensional functional of a hypothetical local density and its external-field couplings that is rich enough to discriminate reactive from inert environments. Formally, the map

$$(p, t) \longmapsto \mathrm{Ep}_p(t)$$

from configuration space (and, in principle, time) to environment space induces an embedded environment–vector manifold with a well-defined metric and a sub-Gaussian reference measure, so that distances, expectations, and kernels in environment space are mathematically well posed.

On top of this representation we provide a fully deterministic, metal-aware implementation for structural snapshots. Given a PDB or mmCIF file, our pipeline constructs for each residue a fixed-dimensional environment vector by combining density-inspired terms from a promolecular surrogate density, metal- and phosphate-like field components obtained by integrating over atom-centred kernels, solvent-accessible surface area, curvature-based quantities, and simple neighbour-geometry summaries, optionally augmented by residue-identity embeddings. The implementation depends only on standard scientific Python libraries and contains no data-driven fitting stage, so that identical inputs produce identical environment vectors under a fixed software environment.

The core question of this paper is not whether a particular machine-learning architecture can achieve state-of-the-art performance, but rather *how much functional structure is already exposed* by this environment representation. To probe this, we treat the environment vectors as features and use an off-the-shelf Random Forest classifier as a shallow, untuned probe. For each protein we curate Tier-1 and Tier-2 sets of functional residues—including catalytic,

metal-coordination, allosteric, and protein–protein interface sites—and evaluate residue-level separation against a background of non-annotated positions. Because many functional sites appear in multiple symmetry-related copies, we employ a site-grouped cross-validation protocol based on `GroupKFold`, in which groups are defined at the level of residue numbers and structural roles, so that no copy of a functional site appears in both training and test folds. The resulting AUC-ROC values characterise how linearly or weakly non-linearly separable the curated functional sets are in environment space, independently of any heavy model tuning.

Conceptually, this work makes three contributions:

(1) **A geometric and measure-theoretic formalisation of residue environments.** We model the collection of residue-wise environment vectors as an embedded manifold equipped with an induced metric and a sub-Gaussian reference measure. This provides a common representation space in which different scoring functions or, in future, learning algorithms can operate, and clarifies how Euclidean, density-overlap, and geodesic kernels on environment space are related.

(2) **A deterministic, metal-aware implementation for structural configurations.** We specify a single, reproducible pipeline that maps arbitrary PDB/mmCIF inputs to fixed-dimensional environment vectors using only structural information and element types. The design is fully deterministic and free of stochastic training, so that identical inputs produce identical outputs under a fixed software environment.

(3) **A unified evaluation protocol that probes the representation rather than a specific model.** We introduce a site-grouped cross-validation scheme and a shallow Random Forest probe to quantify how much functional signal the environment representation exposes across diverse proteins, including large multimeric enzymes and complexes such as Rubisco, GroEL/GroES, SecA, and p53–DNA. The classifier is deliberately kept simple and untuned so that the measured separation reflects properties of the representation itself rather than optimisation of a particular architecture.

Beyond the static case studies considered here, the manifold perspective is deliberately chosen so that both the underlying configuration manifold and the induced environment–vector manifold can serve as targets for future geometric and topological analysis. Time-resolved trajectories $(p, t) \mapsto \mathrm{Ep}_p(t)$ define curves and ensembles in environment space on which persistent homology and related tools can be applied, while interface regions in complexes can be modelled as contact submanifolds equipped with geometric and topological invariants. In this sense, the present residue-level benchmarks are intended as a first validation step within a more general programme: using a single, interpretable environment representation as a common coordinate system for describing intra-protein organisation, inter-protein interactions, and, ultimately, more general molecular systems in $\mathbb{R}^3 \times T$. The remainder of this article develops the mathematical framework, details the construction of the environment vector and the resulting deterministic pipeline, and applies the unified evaluation protocol to a set of representative case studies, with the aim of making the representation and its implementation precise and transparent enough to serve as a reusable front end for subsequent empirical work.

2 Theoretical Framework

2.1 Manifold Representation of Protein Structure

We model a protein's Cα trace as a **1-D curve** $\gamma \subset \mathbb{R}^3 \times T$ and endow it with an induced tubular surface metric. For convenience, we refer to this surface approximation as a "2-D manifold," on which all geometric quantities are evaluated, while the implementation itself operates directly on the Cα point cloud.

The manifold M inherits its metric g from the ambient Euclidean space, making distances and curvatures well-defined. Each residue $p \in M$ is assigned a coordinate neighbourhood U_p comprising all residues within a cutoff radius (typically 5–7 Å). This manifold perspective enables the direct use of geometric concepts—geodesics, principal curvature, and so forth—while treating the protein as a continuous object. As a result, localised analyses become less sensitive to global conformational changes.

2.2 Environment Vector E_p

For each residue p in the manifold, we define an **environment vector** $E_p(t)$ that encodes the local physico-chemical environment around p at a (potential) reaction coordinate or time t:[19, 20]. Formally,

$$E_p(t) = \Big[\frac{1}{\mu(U_p(t))} \int_{U_p(t)} q(x,t)\, \mathrm{dvol}_{g(t)}(x),\ \Big\| \int_{U_p(t)} q(x,t)\ \exp_p^{-1}(x)\, \mathrm{dvol}_{g(t)}(x) \Big\|, \\ \int_{U_p(t)} \kappa(x,t)\, \rho(x,t)\, \mathrm{dvol}_{g(t)}(x),\ \mu(t)\, E_{\text{field}}(t),\ \mathrm{SASA}_{\mathrm{QM}}(p,t),\ \mathrm{Embed}(\mathrm{Residue}(p)) \Big] , \tag{1}$$

where each component (separated by commas) corresponds to a distinct local feature[6, 21]. see supplementary algorithm 1 for code. In this formulation: **(1)** the average partial charge in the neighborhood U_p; **(2)** the norm of the charge-weighted dipole vector emanating from p (i.e. the magnitude of the first moment of charge distribution in U_p); **(3)** a curvature-weighted local density (curvature κ times local mass or atomic density ρ integrated over U_p); **(4)** an external field term capturing any uniform external influence (scaled by a time-dependent factor $\mu(t)$, which could represent, for example, a time-varying field or global conformational change); **(5)** a solvent-accessible surface area (SASA) term for residue p, quantifying how exposed the local environment of p is to the surrounding solvent; conceptually, we view this as an "electron-density SASA" $\mathrm{SASA}_{\mathrm{QM}}(p)$, defined as the area of an isodensity surface patch $\Sigma_{\rho_0} = \{x \in \mathbb{R}^3 \mid \rho(x) = \rho_0\}$ associated with residue p for a suitable threshold ρ_0, so that $\mathrm{SASA}_{\mathrm{QM}}$ is a functional of the underlying electron density ρ, whereas in the present work we approximate $\mathrm{SASA}_{\mathrm{QM}}(p)$ by the standard classical SASA computed from atomic coordinates and van der Waals radii using the Shrake–Rupley algorithm with a probe sphere of radius 1.4 Å, and we interpret the resulting SASA component in E_p as a geometric surrogate of this density-defined exposure functional; and **(6)** an encoding of the residue's identity (in the present setting, each amino acid is embedded into a 4-dimensional vector whose components are four relative physicochemical descriptors). The optional time parameter t allows modeling of reactive events or conformational changes as a smooth path on the manifold (e.g., a reaction coordinate or trajectory frame index).

In practice, we implement a slight extension of this theoretical vector. We denote by $E_p^{\text{theory}}(t)$ the 9-component vector defined above, which captures the core geometric and chemical features. For computational experiments and predictive modeling, we augment this with additional empirically useful features to form a 19-dimensional **simulation vector** $E_p^{\text{sim}}(t)$. The extra geometric components extend $E_p^{\text{theory}}(t)$ with a 10-dimensional block of neighbor-based descriptors, grouped as follows. **(7)**–**(9)** form a two-shell radial density profile: **(7)** inner neighbor count $N_{\text{in}}(p)$, **(8)** outer neighbor count $N_{\text{out}}(p)$, and **(9)** shell ratio $R_{\text{shell}}(p) = N_{\text{in}}(p)/(N_{\text{out}}(p) + \varepsilon)$. **(10)**–**(11)** are radial distribution moments of the neighbor distances $r_q = d(p, q)$ for $q \in U_p$: **(10)** mean neighbor distance $\bar{r}(p)$ and **(11)** radial variance $\text{Var}_r(p)$. **(12)**–**(14)** are angular shape descriptors derived from the eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$ of the local neighbor-direction tensor T_p: **(12)** linearity$(p) = \lambda_1 - \lambda_2$, **(13)** planarity$(p) = \lambda_2 - \lambda_3$, and **(14)** isotropy$(p) = \lambda_3$. Finally, **(15)**–**(16)** provide a local surface-curvature surrogate: **(15)** mean curvature H_p and **(16)** Gaussian curvature K_p, obtained by fitting a quadratic patch to the neighbor cloud in a local tangent frame. Collectively, these ten geometric components form a low-dimensional summary of radial and angular moments of the local environment measure around p; they do not introduce new physical assumptions beyond the manifold–density formalism of E_p^{theory}, but in practice improve robustness on noisy, discretized structures. We emphasize that E_p^{theory} is conceptually a subset of E_p^{sim}; the added components involve no new fundamental assumptions, only additional data-driven inputs. All features are normalized to dimensionless units and standardized (zero-mean, unit-variance) before statistical analysis to prevent scale disparities. We view Ep as a coarse-grained functional of the underlying electron density and geometry: many distinct densities can map to the same Ep, and our goal is not to reconstruct $\rho(x, t)$ but to retain enough information to separate reactive from non-reactive residues.

Formal definition of the environment vector. A rigorous metric–probabilistic formalism of the environment–vector space (E,d) is provided in Appendix E. In brief, we treat the map $\iota : M \times T \hookrightarrow \mathbb{R}^k,\ p \mapsto E_p$ as a C^1 immersion whose image $E = \iota(M \times T)$ is an r-dimensional embedded sub-manifold of $\mathbb{R}^k$ (Definition E.1 in Appendix E). The Euclidean metric, the embedded L^2 metric, and the density–overlap metric all induce the same topology and render (E, d) complete (Theorem E.2, Appendix E).

2.3 Fiber bundle for reactive states

To incorporate site-specific chemical information, we attach a fiber to the manifold at each residue. Formally, we introduce a fiber bundle $\pi : E \to M$ whose fiber $F_p = \pi^{-1}(p)$ represents the **reactive state** of residue p. In regions that are not reactive, F_p is trivial (carrying no additional data). At a putative reactive site, however, F_p contains a localised quantum state—such as an electron-density distribution or wavefunction—that encodes that residue's reactivity potential. These fiber states can be thought of as small "quantum attachments" to the geometric manifold: they add high-resolution chemical detail exactly where it matters (e.g. at an active site) while leaving inert regions described solely by the coarse geometry.

In practical terms, a non-trivial fiber can be instantiated by a pre-computed electron-density cloud or an analytic function (Gaussian, etc.) centred on that residue. For example, if a residue is known to coordinate a metal or participate in catalysis, one might attach a

localised orbital or electron-density patch representing that functionality. By contrast, all other residues carry the trivial fiber (identity map), meaning they are treated uniformly (e.g. assigned a generic chemical identity such as carbon with no special reactivity). The transition between trivial and non-trivial fibres is defined smoothly to avoid discontinuities on M. This construction draws on quantum chemistry—via density-functional or semi-empirical calculations—to supply the fibre data for reactive sites.

The fiber-bundle framework thus creates a two-scale model: the base manifold M captures global and local geometry for all residues, whereas the fiber F_p adds quantum-level detail for specific key sites. The benefit is an efficient focusing of computational effort: highly detailed physics is included only in localised regions (e.g. active sites), whereas the rest of the protein is treated in a simplified manner. In essence, the protein is represented as an **augmented manifold** where only a subset of points carry extra chemical degrees of freedom.

When the localised fibres of two residues p, q intersect, they generate an *interaction submanifold*

$$\mathcal{M}_{pq} = \big\{\, x \in \mathbb{M} \mid \psi_p(x) \neq 0,\ \psi_q(x) \neq 0 \big\},$$

on which we propose to measure purely geometric invariants such as the area $\mathrm{Area}(\mathcal{M}_{pq})$, Euler characteristic $\chi(\mathcal{M}_{pq})$, and fibre-overlap homology classes. Here, $\mathbb{M} = T \times M$ denotes space–time, with $x = (t, \mathbf{r})$ capturing both temporal and spatial coordinates. These invariants—rather than a single scalar score—could form the basis of our deductive protein–protein interface analysis. In the present work, all numerical results are obtained solely from the base manifold and the environment vectors Ep; non-trivial fibers and their overlaps are introduced here as a theoretical extension for future interface-level analysis.

3 Methods

3.1 Overview of the 1–9 Environment Pipeline

We implemented a full residue-centric environment pipeline that maps a static protein structure to per-residue environment vectors used for downstream scoring. The pipeline consists of nine stages: (1) structure parsing and residue/atom indexing, (2) residue–residue contact graph construction, (3) surrogate (promolecular) electron density generation, (4) optional quantum-corrected density replacement on a selected subset of residues, (5) density-derived local descriptors, (6) charge-derived descriptors (quantum or surrogate), (7) definition of local integration domains U_p, (8) computation of theoretical environment components, and (9) assembly of final environment vectors. All stages are executed in a single integrated script to guarantee consistent fragment definition and parameter sharing.

3.1.1 Input Structures and Residue Representation (Step 1)

Protein structures were read from PDB or mmCIF files using standard parsers. Only the first structural model was used. Each residue p was represented by a dictionary containing chain identifier, residue sequence number, optional insertion code, residue type, hetero-flag, and a list of atoms. Each atom was assigned a global integer index and stored with element

type and Cartesian coordinates $\mathbf{r}_a \in \mathbb{R}^3$. Residues with zero atoms were ignored. The residue identifier was formatted as `chain:resseq[icode]` (e.g., `A:201`).

3.1.2 Solvent Accessible Surface Area (SASA) (Step 1b)

Per-residue solvent accessible surface area was computed by the Shrake–Rupley algorithm as implemented in MDTraj (probe radius 1.4Å, residue mode). If MDTraj was unavailable or if residue mapping failed, SASA values were set to zero. Residue-level mapping was performed by matching $(\text{chain}, \text{resSeq}, \text{icode})$ keys between MDTraj topology and the parsed residue list. Let $\text{SASA}(p)$ denote the resulting SASA for residue p. We also computed the mean SASA over residues, $\overline{\text{SASA}}$, for later scaling.

3.1.3 Residue Contact Graph and Fragment Definition (Step 2)

To define local neighborhoods, we computed residue–residue contacts using atomic proximity. Let r be a fixed chart radius. For every atom a we queried all atoms b satisfying

$$\|\mathbf{r}_a - \mathbf{r}_b\| < 2r.$$

These queries were performed using a global KDTree over all atomic coordinates; when KDTree was unavailable, an $O(N^2)$ fallback was used. If any atom pair across two distinct residues satisfied the condition, the residues were considered overlapping neighbors. This generated (i) a residue adjacency list $\mathcal{N}(p)$ ("overlaps") and (ii) atom-level contact pairs for diagnostic output.

For each residue p, we defined its *fragment* F_p as the atom set of p union all atoms from neighboring residues:

$$F_p = \bigcup_{q \in \{p\} \cup \mathcal{N}(p)} \text{atoms}(q).$$

All subsequent density and charge calculations for residue p were performed on F_p. This ensures that the same neighborhood definition is used consistently throughout Steps 3–9.

3.1.4 Surrogate Promolecular Electron Density (B-mode) (Step 3)

Because we do not assume access to full quantum electron densities for large proteins, we constructed a local surrogate density on a 3D grid around F_p. For each fragment we built a Cartesian grid $\{x_i\}$ spanning the fragment bounding box with an isotropic padding P (default 2.0Å) to avoid truncating Gaussian tails. The grid spacing was $\Delta = 0.5$Å. Let $\Delta V = \Delta^3$. Although the theoretical formulation defines Up(t) as a small geodesic neighborhood around residue p, in practice we approximate Up(t) by the full fragment grid Fp generated from residue p and its spatial neighbors. This approximation provides stable numerical integration and avoids under-sampling issues on narrow local grids.

We defined the B-mode promolecular density as a sum of element-wise Gaussian atomic kernels:

$$\rho_B(x) = \sum_{a \in F_p} Z_a \exp\left(-\frac{\|x - \mathbf{r}_a\|^2}{2\sigma_a^2}\right).$$

Here Z_a is a rough nuclear charge proxy taken from a fixed lookup table, and σ_a is a heuristic width chosen on the scale of atomic radii. This density is not intended to reproduce quantum densities; it is a smooth surrogate field for residue-local integration.

3.1.5 Optional Quantum-Corrected Density (A-mode) (Step 4)

For a user-selected residue subset $\mathcal{A}$, we computed semiempirical charges using GFN2-xTB on the corresponding fragment F_p. Each fragment was exported to an XYZ file and xTB was executed with the `--charges` option to obtain per-atom partial charges $\{q_a^{\mathrm{xTB}}\}_{a\in F_p}$. If xTB failed, the residue reverted to B-mode.

When charges were obtained, we produced an A-mode corrected density by reweighting the same Gaussian kernels:

$$\rho_A(x) = \sum_{a\in F_p} \max(Z_a - q_a^{\mathrm{xTB}}, \epsilon)\, \exp\left(-\frac{\|x-\mathbf{r}_a\|^2}{2\sigma_a^2}\right),$$

with $\epsilon = 0.1$ to maintain positivity. For residues not in $\mathcal{A}$, or for failed xTB runs, the final density for feature extraction was ρ_B. We denote the final density used for residue p by $\rho_p \in \{\rho_B, \rho_A\}$.

3.1.6 Density-Derived Local Descriptors and Element-Resolved Channels (Step 5)

Using the final density ρ_p on its fragment grid, we computed a set of density-based descriptors around residue p.

Let r_p be the geometric center of residue p (mean of its atomic coordinates). The total local density was approximated by a Riemann sum over the grid points $\{x_i\}$ associated with fragment F_p:

$$D_p \;=\; \int_{U_p} \rho_p(x)\, dx \;\approx\; \sum_{i\in U_p} \rho_p(x_i)\, \Delta V,$$

where $\Delta V = \Delta^3$ is the grid cell volume and U_p is the local integration mask defined in Section 1.9.

We defined a normalized density barycenter c_p as

$$c_p \;\approx\; \frac{1}{D_p} \sum_{i\in U_p} \rho_p(x_i)\, x_i\, \Delta V,$$

provided that $D_p > 0$; otherwise c_p was set to r_p.

We also computed a dipole-like density moment relative to r_p,

$$d_p^{\rho} \;\approx\; \sum_{i\in U_p} \rho_p(x_i)\, \big(x_i - r_p\big)\, \Delta V,$$

which serves as a coarse descriptor of anisotropy in the local density distribution.

Finally, we computed a curvature-like smoothness scalar from the density gradient:

$$C_p^\rho \;=\; \int_{U_p} \|\nabla \rho_p(x)\|^2 \, dx \;\approx\; \sum_{i \in U_p} \|\nabla \rho_p(x_i)\|^2 \, \Delta V,$$

where $\nabla \rho_p$ was evaluated by finite differences on the fragment grid under implicit Neumann boundary conditions. This term is a surrogate for a "curvature-weighted density" and should be interpreted as a smoothness/roughness measure of the surrogate density rather than a geometric curvature of a fixed isosurface.

Element-resolved density channels. To capture the influence of chemically distinct atom groups without using any residue-type labels, we additionally decomposed ρ_p into element-resolved channels. Let $\mathcal{A}_p$ denote the atom set of fragment F_p, and define disjoint subsets based on element type and hetero flags:

- $\mathcal{A}_p^{(\mathrm{prot})}$: protein heavy atoms (standard amino-acid residues),
- $\mathcal{A}_p^{(\mathrm{metal})}$: metal atoms (e.g. Mg, Mn, Zn, Ca, Fe, Ni, Cu, Co),
- $\mathcal{A}_p^{(\mathrm{P})}$: phosphorus atoms,
- $\mathcal{A}_p^{(\mathrm{hal})}$: halogen atoms (F, Cl, Br, I),
- $\mathcal{A}_p^{(\mathrm{het})}$: non-protein heavy atoms from HETATM records (excluding crystallographic waters).

For each channel $c \in \{\mathrm{prot}, \mathrm{metal}, \mathrm{P}, \mathrm{hal}, \mathrm{het}\}$ we defined a channel-specific density by restricting the Gaussian sum in ρ_p to the corresponding atom subset:

$$\rho_p^{(c)}(x) \;:=\; \sum_{a \in \mathcal{A}_p^{(c)}} w_a \exp\!\left(-\frac{\|x - r_a\|^2}{2\sigma_a^2}\right),$$

with the same kernel widths σ_a as in Section 1.5 and weights w_a inherited from the B-mode/A-mode definition (either Z_a or the charge-reweighted coefficient).

We then computed a channel-resolved local density for each c:

$$D_p^{(c)} \;\approx\; \sum_{i \in U_p} \rho_p^{(c)}(x_i) \, \Delta V,$$

and, when $D_p > 0$, the corresponding fractional contributions

$$f_p^{(c)} \;=\; \frac{D_p^{(c)}}{D_p}.$$

The scalars $\{D_p^{(c)}, f_p^{(c)}\}_c$ enter the environment vector as additional components. They quantify how much of the local surrogate density around residue p is carried by protein atoms, metals, phosphorus, halogens, or hetero-ligand atoms, without relying on residue-type or binding-site annotations.

3.1.7 Charge-Derived Descriptors (Step 6)

A-mode charges. For residues $p \in \mathcal{A}$ with successful xTB runs, we computed per-residue net charge

$$Q_p^{\mathrm{xTB}} = \sum_{a \in p} q_a^{\mathrm{xTB}},$$

and a charge dipole moment

$$\mathbf{d}_p^q = \sum_{a \in p} q_a^{\mathrm{xTB}} (\mathbf{r}_a - r_p).$$

These quantities were computed using the same fragment F_p and atom indexing as above to prevent mismatch.

Residue-level charge surrogates (B-mode) In the theoretical environment vector, the charge-like component is intended to represent a coarse local average of the underlying charge field $q(x)$ over the fragment domain U_p associated with residue p. When quantum mechanical (QM) results are available, we use the xTB-derived residue net charges Q_p^{xTB} and their associated dipole moments as in Section 1.5. For structures without QM data, we construct a residue-level surrogate charge purely from atomic element information, without using any residue-type labels.

Atom-based base charge Q_p^{atom}. Let $\mathcal{A}_p$ denote the set of atoms belonging to residue p, and let E_a be the element symbol of atom $a \in \mathcal{A}_p$ (e.g. C, N, O, S, Mg). We define a simple element-dependent weight function $w_{\mathrm{elem}} : \{\text{elements}\} \to \mathbb{R}$ that encodes coarse electrostatic tendencies:

$$w_{\mathrm{elem}}(E) \;=\; \begin{cases} -0.40, & E \in \{\mathrm{O}\}, \\ +0.30, & E \in \{\mathrm{N}\}, \\ -0.25, & E \in \{\mathrm{S}, \mathrm{SE}\}, \\ -0.35, & E \in \{\mathrm{P}\}, \\ -0.20, & E \in \{\mathrm{F}, \mathrm{CL}, \mathrm{BR}, \mathrm{I}\}, \\ +0.50, & E \in \{\mathrm{MG}, \mathrm{CA}\}, \\ +0.60, & E \in \{\mathrm{MN}, \mathrm{FE}, \mathrm{ZN}, \mathrm{CU}, \mathrm{NI}, \mathrm{CO}\}, \\ 0.0, & \text{otherwise.} \end{cases}$$

This mapping is not meant to approximate formal charges or partial charges; instead, it is a coarse, element-resolved surrogate that assigns negative weights to electronegative atoms (O, halogens, P, S/Se) and positive weights to typical cationic metals. Carbon and hydrogen are treated as neutral background and contribute 0 in this B-mode surrogate.

The atom-based base charge for residue p is then defined as

$$Q_p^{\mathrm{atom}} \;=\; \sum_{a \in \mathcal{A}_p} w_{\mathrm{elem}}(E_a).$$

This quantity depends only on the atomic composition of the residue and does not use any residue-type or motif annotation.

SASA-modulated B-mode charge Q_p^{B}. As in the original B-mode construction, we modulate the base charge by a SASA-dependent scaling factor to reflect the fact that buried charged groups are less exposed to the solvent and external field than solvent-accessible ones.

Let $\mathrm{SASA}(p)$ denote the solvent-accessible surface area of residue p, and let $\overline{\mathrm{SASA}}$ be the mean SASA across all residues in the structure (Section 1.3). We define a scaling factor

$$s_p \;=\; \phi\left(\frac{\mathrm{SASA}(p) - \overline{\mathrm{SASA}}}{\overline{\mathrm{SASA}}}\right),$$

where ϕ is a bounded, monotone function such as

$$\phi(z) \;=\; \mathrm{clip}\big(1 + \alpha z, s_{\min}, s_{\max}\big),$$

with a user-specified slope α and lower/upper caps $s_{\min}$, $s_{\max}$ (e.g. $s_{\min} = 0.5$, $s_{\max} = 1.5$). In the implementation, this scaling corresponds to the function `env_scale_from_sasa` applied to the residue-level SASA.

The SASA-modulated B-mode charge is

$$Q_p^{\mathrm{B}} \;=\; s_p\, Q_p^{\mathrm{atom}}.$$

Charge component in the environment vector. For residues with xTB-derived charges available, we use

$$q_p \;=\; Q_p^{\mathrm{xTB}}$$

as the charge component in the theoretical environment vector. For all other residues, we fall back to the atom-based surrogate,

$$q_p \;=\; Q_p^{\mathrm{B}}.$$

This hierarchy ensures that the theoretical environment vector remains free of any residue-type or motif-based charge assignment. The only inputs are atomic positions, element types, and SASA, which are consistently available in static PDB/mmCIF structures.

3.1.8 Local Integration Domains U_p (Step 7)

The theoretical model defines $U_p(t)$ as an open neighborhood around residue p. In implementation, U_p is induced by the fragment grid and residue atoms. Specifically, let R_{env} be an environment-ball radius (default 2.0Å). We defined

$$U_p = \bigcup_{a \in p} B(\mathbf{r}_a, R_{\mathrm{env}}) \cap \{x_i\},$$

i.e., the set of grid points within R_{env} of any atom in residue p. All theoretical integrals over U_p were approximated by sums over these masked grid points.

3.1.9 Theoretical Environment Components (Step 8)

We assembled a theoretical environment vector E_p^{theory} corresponding to Eq. (1) of the model. Continuous fields $q(x)$, $\rho(x)$, and $\kappa(x)$ are represented discretely: $q(x)$ is taken as a residue-level net charge (quantum or surrogate), and $\rho(x)$ is the surrogate density ρ_p defined in Sections 1.5–1.6. The components are:

1. Average charge. We used residue net charge as a coarse approximation to $\mu(U_p)^{-1} \int_{U_p} q(x)\, dx$:

$$q_p \;=\; \begin{cases} Q_p^{\text{xTB}} & \text{if xTB charges are available,} \\ Q_p^{\text{B}} & \text{otherwise.} \end{cases}$$

2. Dipole magnitude. We used $\|d_p^q\|$ if xTB charges were available, and otherwise fell back to the density-dipole magnitude $\|d_p^\rho\|$ from Section 1.7.

3. Curvature-weighted density surrogate. We set this term to C_p^ρ, acknowledging that it is a smoothness proxy derived from ρ_p rather than a geometric curvature of a fixed isosurface.

4. External-field block from element-resolved sources (two variants). To capture the influence of chemically distinct external sources (e.g. metals, phosphorus-rich cofactors, or halogenated ligands) without using biochemical labels, we define a compact three-dimensional external-field block for each residue p,

$$\text{external_field}_p := \left\| F_p^{(\text{metal})},\; F_p^{(P)},\; F_p^{(\text{hal})} \right\|,$$

which then enters the environment vector as a single structured component.

We implemented two closely related constructions of $F_p^{(c)}$, $c \in \{\text{metal}, P, \text{hal}\}$:

(i) Global Gaussian external field (structure-wide variant). In the first variant, we define smooth scalar field descriptors at the residue center r_p by summing distance-weighted contributions from all atoms of the relevant element class in the entire structure. Let $\mathcal{A}^{(\text{metal})}$, $\mathcal{A}^{(P)}$, and $\mathcal{A}^{(\text{hal})}$ denote the global sets of atoms in the structure whose element types belong to user-specified metal, phosphorus, and halogen lists, respectively (e.g. metals: Mg, Mn, Zn, Ca, Fe, Ni, Cu, Co; halogens: F, Cl, Br, I). For each residue p and each channel $c \in \{\text{metal}, P, \text{hal}\}$, we set

$$F_p^{(c)} \;=\; \sum_{b \in \mathcal{A}^{(c)}} w_b \exp\left(-\frac{\left\| r_p - r_b \right\|^2}{2\tau_c^2} \right),$$

where w_b is either the atomic number Z_b or a constant weight, and τ_c is a fixed length scale (on the order of a few Å) controlling how rapidly the contribution decays with distance. This variant yields a global, structure-wide external field evaluated at r_p and requires an additional loop over the corresponding atom sets.

(ii) Channel-based local external field (density-channel variant). In the second variant, we reuse the element-resolved density channels already computed in Step 5. Recall that for each fragment F_p we defined channel-specific densities $\rho_p^{(c)}(x)$ and their integrated local contributions

$$D_p^{(c)} \;\approx\; \sum_{i \in U_p} \rho_p^{(c)}(x_i)\, \Delta V,$$

together with the total local density

$$D_p \;=\; \sum_{i \in U_p} \rho_p(x_i)\,\Delta V$$

and the fractional contributions

$$f_p^{(c)} \;=\; \frac{D_p^{(c)}}{D_p} \quad \text{when } D_p > 0 \quad \text{(Section 1.7)}.$$

In the channel-based variant, we simply set

$$F_p^{(c)} := f_p^{(c)} \quad \text{for } c \in \{\text{metal}, P, \text{hal}\},$$

with $F_p^{(c)} = 0$ if $D_p = 0$. Thus,

$$\text{external_field}_p := \left\| F_p^{(\text{metal})},\; F_p^{(P)},\; F_p^{(\text{hal})} \right\|$$

is a compact, element-resolved external-field surrogate derived purely from the local density channels around residue p. This construction does not require any additional grid passes or global atom loops beyond those already used to compute the density channels, so the asymptotic cost of Step 8 is unchanged.

Both variants depend only on atomic positions and element types and remain free of any residue-type or binding-site annotations. The global Gaussian variant (i) explicitly aggregates contributions from all metals, phosphorus atoms, and halogens in the structure and is useful when one wishes to treat these as long-range field sources. In contrast, the channel-based variant (ii) treats the external-field block as a normalized summary of the local density composition around residue p, which is more computationally lightweight and more tightly coupled to the fragment F_p.

In all experiments reported in this work we used the channel-based variant (ii) as the default choice for external_field$_p$, while keeping the global Gaussian variant (i) available in the reference implementation for ablation and sensitivity studies.

5. SASA. We inserted SASA(p) as a scalar descriptor of local solvent exposure (Section 1.3).

6. Residue embedding. Each residue type was optionally encoded as a four-dimensional physicochemical vector as described in Section 1.11. In ablation experiments focused on purely geometric and element-based descriptors, this embedding can be dropped.

The concatenation of these components defines the final theoretical environment vector E_p^{theory} used in downstream analysis and supervised evaluation.

3.1.10 Physicochemical 4D Residue Embedding (Step 8, component 6)

To provide a low-dimensional but interpretable residue identity term, we mapped each residue to a 4D real vector

$$\text{Emb}(p) = [q_{\text{formal}},\, h_{\text{hydropathy}},\, s_{\text{steric}},\, a_{\text{aromatic}}] \in \mathbb{R}^4.$$

The coordinates follow a physically motivated ordering:

1. q_{formal}: formal side-chain charge at near-neutral pH (e.g., ASP/GLU = -1, LYS/ARG = $+1$, others 0).

2. $h_{\text{hydropathy}}$: normalized hydropathy index (Kyte–Doolittle scale, linearly rescaled to $[-1, 1]$).

3. s_{steric}: normalized steric size/volume proxy (linearly rescaled to $[0, 1]$).

4. a_{aromatic}: aromaticity indicator (1 for aromatic residues, 0 otherwise).

This embedding replaces one-hot encodings and is designed to preserve coarse physicochemical similarity while remaining lightweight for inference.

3.1.11 Geometric Neighbor Augmentation (Step 9 extra features)

In addition to E_p^{theory}, we augment each residue p with a compact family of purely geometric descriptors computed from the contact graph $\mathcal{N}(p)$ of Cα-based residue centers. These descriptors are interpreted as low-order radial and angular moments of the local environment around p:

1. **Two-shell radial density profile.** For a fixed cutoff radius $r_{\max}$, we split the neighborhood into an inner and an outer shell,

   $$\mathcal{N}_{\text{in}}(p) = \{q \in \mathcal{N}(p) : d(p, q) \leq r_{\max}/2\}, \qquad \mathcal{N}_{\text{out}}(p) = \{q \in \mathcal{N}(p) : r_{\max}/2 < d(p, q) \leq r_{\max}\}.$$

   We record the inner and outer counts

   $$N_{\text{in}}(p) = |\mathcal{N}_{\text{in}}(p)|, \qquad N_{\text{out}}(p) = |\mathcal{N}_{\text{out}}(p)|,$$

   together with the shell ratio

   $$R_{\text{shell}}(p) = \frac{N_{\text{in}}(p)}{N_{\text{out}}(p) + \varepsilon},$$

   which distinguishes compact, core-like neighborhoods from more diffuse or one-sided environments.

2. **Radial distribution moments.** Let $r_q = d(p, q)$ for $q \in \mathcal{N}(p)$. We compute the first two moments of the radial distribution,

   $$\bar{r}(p) = \frac{1}{|\mathcal{N}(p)|} \sum_{q \in \mathcal{N}(p)} r_q, \qquad \text{Var}_r(p) = \frac{1}{|\mathcal{N}(p)|} \sum_{q \in \mathcal{N}(p)} \big(r_q - \bar{r}(p)\big)^2,$$

   capturing both the typical neighbor distance and how tightly neighbors are concentrated around that radius.

3. **Angular shape tensor and anisotropy.** For each neighbor we define a unit direction vector

$$u_q = \frac{x_q - x_p}{\|x_q - x_p\|}, \qquad q \in \mathcal{N}(p),$$

and form the 3×3 direction tensor

$$T_p = \frac{1}{|\mathcal{N}(p)|} \sum_{q \in \mathcal{N}(p)} u_q u_q^\top.$$

Let $\lambda_1 \geq \lambda_2 \geq \lambda_3$ denote the eigenvalues of T_p. From these we derive scalar shape descriptors

$$\text{linearity}(p) = \lambda_1 - \lambda_2, \qquad \text{planarity}(p) = \lambda_2 - \lambda_3, \qquad \text{isotropy}(p) = \lambda_3,$$

which distinguish bulk-like, planar-surface, and ridge-like local packing around p.

4. **Local surface curvature surrogate.** In the local coordinate frame defined by the principal directions of T_p, we fit a quadratic height function

$$z = ax^2 + by^2 + cxy$$

to the neighbor cloud by least squares. The fitted coefficients are converted into mean and Gaussian curvatures (H_p, K_p), providing a coarse estimate of whether p lies in a convex bulge, concave pocket, or saddle-like region.

Residue centers were defined as geometric means of atomic coordinates.

3.1.12 Final Environment Vectors (Step 9)

The final theoretical environment vector for residue p collects the components defined in Section 1.10, including the external-field block

$$\text{external_field}_p := \left\| F_p^{(\text{metal})},\ F_p^{(P)},\ F_p^{(\text{hal})} \right\|.$$

We write

$$E_p^{\text{theory}} = \big(q_p,\ \|d_p\|,\ C_{\rho,p},\ \text{external_field}_p,\ \text{SASA}(p),\ \text{Emb}(p)\big).$$

Here q_p is the residue-level charge component (quantum or surrogate), $\|d_p\|$ is the charge- or density-derived dipole magnitude, C_p^ρ is the curvature-weighted density surrogate, $\big(F_p^{(\text{metal})}, F_p^{(P)}, F_p^{(\text{hal})}\big)$ is the external-field block defined from element-resolved sources, $\text{SASA}(p)$ is the solvent-accessible surface area, and $\text{Emb}(p)$ is the optional 4D physicochemical residue embedding.

For predictive experiments we augment E_p^{theory} with ten purely geometric neighbor descriptors from Section 1.12 to obtain the simulation vector

$$E_p^{\text{sim}} = \big[E_p^{\text{theory}},\ N_p^{\text{in}},\ N_p^{\text{out}},\ R_p^{\text{shell}},\ \bar{r}_p,\ \text{Var}(r)_p,\ \lambda_p^{(1)},\ \lambda_p^{(2)},\ \lambda_p^{(3)},\ H_p,\ K_p\big].$$

where N_p^{in}, N_p^{out}, R_p^{shell}, $\bar{r}_p$, $\text{Var}(r)_p$, $\lambda_p^{(1)}$, $\lambda_p^{(2)}$, $\lambda_p^{(3)}$, H_p, and K_p are the neighbor-count, radial-moment, anisotropy, and curvature descriptors defined in Section 3.1.11. Stacking over residues yields the matrices saved as `Ep_theory.npy` and `Ep_sim.npy`, with a manifest file recording residue ordering, feature names, and parameter settings.

3.1.13 Outputs and Reproducibility

The pipeline produces: `meta.json` (residues, overlaps, SASA, atom contacts), `density_final.json` (A/B density provenance), optional density grids (`rhoB_*.npy`, `rhoA_*.npy`), `env_features.json` (Steps 5–6 descriptors), `Ep_theory.npy` and `Ep_sim.npy` with `Ep_manifest.json`. All parameters (r, P, Δ, R_{env}, xTB options, SASA scaling α) are explicitly stored in the manifest to ensure exact reproduction.

3.2 Supervised evaluation of residue-level feature representations

3.2.1 Residue-level feature matrix

For a given protein structure, we represent each residue i by a fixed-length feature vector $\mathbf{x}_i \in \mathbb{R}^d$ that summarizes its local structural and physicochemical environment (e.g. geometric descriptors, solvent accessibility, density-based quantities, or learned embeddings). Stacking all residue-level feature vectors yields a matrix

$$X \in \mathbb{R}^{N \times d},$$

where N is the number of residues in the protein (or in a set of proteins) and d is the number of features per residue.

When features are computed from heterogeneous sources, missing values may arise for specific residues or descriptors. To handle this in a reproducible way, we apply a simple column-wise mean imputation separately within each training split: for each feature dimension $j \in \{1, \ldots, d\}$, any missing entries X_{ij} in the training data are replaced by the empirical mean of the non-missing values in column j computed on the training residues of that split. The fitted imputer is then applied to the corresponding test residues of the same split. In code, this corresponds to using a `SimpleImputer(strategy="mean")` fitted on the training subset only and then applied to both training and test subsets.

3.2.2 Definition of residue-level labels

Each residue i is assigned a binary label $y_i \in \{0, 1\}$ according to an externally curated notion of *functional involvement*, based on structural and biochemical information from the literature (active-site annotations, UniProt/PDB functional annotations, mutational scanning, interface mapping, etc.). Rather than fixing a single notion of "functional" versus "background", we construct a *panel* of binary classification tasks, each defined by a different choice of positive set $\mathcal{P}_t \subseteq \{1, \ldots, N\}$ indexed by t (task index). For a given task t we define

$$y_i^{(t)} = \begin{cases} 1, & \text{if } i \in \mathcal{P}_t, \\ 0, & \text{otherwise}, \end{cases}$$

so that all tasks share the same feature matrix X and evaluation protocol while differing only in the definition of the positive class. This allows us to probe how well a single environment representation separates different *functional types* of residues from background.

Across all complexes considered in this work, we adopt a common abstract taxonomy of residue roles and then instantiate it on a per–protein basis using structural and biochemical information from the literature. At the top level we distinguish two tiers:

- **Tier 1 (functional core + support).** Residues with strong, local evidence for functional involvement. This tier is further subdivided into role classes that are shared across systems:
  - *catalytic/metal core:* residues that directly participate in chemical catalysis (general acid/base, nucleophiles, enediolate stabilisers) or coordinate essential metal ions;
  - *PTM core:* residues that are themselves sites of well-characterised post-translational modification (e.g. Lys carbamylation/acetylation, phosphorylation) when the modification is mechanistically essential;
  - *gates and latches:* loop or tail residues that open or close access to the active site, or that stabilise well-defined open/closed conformations;
  - *interfaces and anchors (PPI / DNA / ligand):* residues that contribute directly to binding energy at protein–protein, protein–DNA or protein–ligand interfaces (e.g. hydrophobic patches, IVL-like anchors, core beta–hairpins at obligate interfaces);
  - *allosteric and assembly-network pivots:* residues that transmit conformational changes or stabilise allosteric networks (e.g. hinge-support positions, inter-subunit salt-bridge clusters) when these roles are defined at residue resolution in the literature.
- **Tier 2 (structural shell).** Residues that form the immediate structural shell around Tier 1 cores and anchors. Typical examples include residues that maintain the geometry, packing or electrostatic balance of a catalytic pocket, interface anchor or hinge, but are not themselves essential hotspots. In practice, Tier 2 residues are defined by explicit, local neighbourhood rules (e.g. fixed residue-index ranges or direct-contact shells around Tier 1 cores) for each protein and may or may not have direct site-specific mutational evidence.

For each protein complex we instantiate this abstract scheme by assigning individual residues to one or more of the Tier 1 role classes based on high-confidence structural and biochemical evidence (e.g. active-site annotations, metal ligands, mutational scanning, interface mapping), and then defining Tier 2 shells by explicit local neighbourhood rules around these cores. Residue lists for each role and tier (e.g. catalytic core vs. PTM core vs. gate vs. interface vs. allosteric shell in a given complex) are defined in a protein-specific manner but always drawn from the above role taxonomy.

Given these assignments, we define several families of label sets:

- *global Tier 1* ($\mathcal{P}_{\mathrm{T1-all}}$): the union of all Tier 1 residues across all role classes in a given complex;
- *global Tier 2* ($\mathcal{P}_{\mathrm{T2-all}}$): the union of all Tier 2 shell residues in that complex;
- *role-specific Tier 1 sets* ($\mathcal{P}_r$): for each role class r (catalytic/metal core, PTM core, gates/latches, interface anchors, allosteric/assembly pivots), the subset of Tier 1 residues assigned to that role;

- *role-specific Tier 1+2 unions* ($\mathcal{P}_{r,\mathrm{T1+T2}}$): for roles where a clear supporting shell is well defined (e.g. catalytic basin, clamp, relay helix), the union of the Tier 1 core and its Tier 2 shell.

Each such set $\mathcal{P}_t$ defines a binary classification task as above, with all remaining residues in the complex treated as negatives. These negatives should be interpreted as unlabeled background rather than as a claim of functional irrelevance for those residues. Importantly, several biologically relevant mechanisms are *intentionally left outside* the label space and thus treated as background in all tasks: residues whose importance is dominated by long-range electrostatic steering, intrinsically disordered segments, phase-separating low-complexity regions or global mechanical properties of large-scale domain motions are not explicitly annotated in the present work. The scope of the current benchmarks is therefore restricted by design to functional types that are expected to leave a strong signature in a static, local environment descriptor. For notational simplicity, we omit the task index t in what follows and write $\mathbf{y} = (y_1, \ldots, y_N)^\top \in \{0,1\}^N$ for a generic label vector; all procedures are applied independently to each label set in the panel.

3.2.3 Random forest classifier and cross-validation protocols

To quantify how well a given feature representation separates functional from background residues under a particular label definition, we train a supervised classifier and evaluate its performance using the area under the receiver operating characteristic curve (AUC-ROC).

We use a random forest classifier (scikit-learn `RandomForestClassifier`) as a non-parametric baseline. Unless stated otherwise, we fix the hyperparameters as follows:

- `n_estimators = 100`: number of trees in the ensemble;
- `class_weight = "balanced"`: automatic class weighting inversely proportional to class frequencies to counter label imbalance;
- `random_state = 42`: fixed random seed to ensure reproducibility of tree construction and sampling;
- `n_jobs = 1`: single-threaded execution to avoid non-determinism from parallel scheduling.

Let X^{imp} denote the imputed feature matrix after applying the mean imputer within each training split, as described above. We consider two related evaluation protocols.

Residue-stratified cross-validation As a baseline, we evaluate performance with stratified K-fold cross-validation at the residue level:

1. We construct a **stratified** split of the indices $\{1, \ldots, N\}$ into K folds (typically $K = 5$), preserving the overall positive:negative ratio in each fold. In code, this corresponds to `StratifiedKFold(n_splits=5, shuffle=True, random_state=42)`.
2. For each fold $k \in \{1, \ldots, K\}$:

(a) Define the training indices $\mathcal{I}_{\text{train}}^{(k)}$ and test indices $\mathcal{I}_{\text{test}}^{(k)}$.

(b) Fit the mean imputer on $X_{\mathcal{I}_{\text{train}}^{(k)}}$ and transform both training and test feature matrices to obtain $X_{\mathcal{I}_{\text{train}}^{(k)}}^{\text{imp}}$ and $X_{\mathcal{I}_{\text{test}}^{(k)}}^{\text{imp}}$.

(c) Fit a random forest classifier on $\left(X_{\mathcal{I}_{\text{train}}^{(k)}}^{\text{imp}}, \mathbf{y}_{\mathcal{I}_{\text{train}}^{(k)}}\right)$.

(d) Obtain predicted class probabilities $\hat{p}_i^{(k)} = \hat{\mathbb{P}}(y_i = 1 \mid \mathbf{x}_i)$ for all $i \in \mathcal{I}_{\text{test}}^{(k)}$ using `predict_proba`.

(e) Compute the AUC-ROC on the k-th test fold, $\text{AUC}^{(k)} = \text{AUC}(\mathbf{y}_{\mathcal{I}_{\text{test}}^{(k)}}, \hat{p}^{(k)})$.

3. We report the mean and standard deviation of the AUC across folds:

$$\overline{\text{AUC}} = \frac{1}{K}\sum_{k=1}^{K} \text{AUC}^{(k)}, \qquad \sigma_{\text{AUC}} = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\text{AUC}^{(k)} - \overline{\text{AUC}}\right)^2}.$$

In this protocol, the basic sampling unit is an individual residue. This setting is appropriate when residues can be treated as approximately independent given the structure, and when symmetric or replicated copies of the same functional site are not included as positives in the same task.

Site-grouped cross-validation to avoid functional-site leakage For symmetric oligomeric complexes (such as chaperonins) or assemblies in which multiple chains share identical functional sites (e.g. repeated ATP-binding pockets or GroES-anchoring motifs), residue-level cross-validation can overestimate performance: different copies of the same functional site may appear in both the training and test sets. To obtain a more conservative estimate of how well the features separate distinct functional sites from background, we additionally perform *site-grouped* cross-validation whenever the label definition includes replicated copies of the same site.

We define a discrete site identifier g_i for each residue i that groups residues belonging to the same functional site across chains. In practice, g_i can be constructed from metadata such as the protein role (e.g. GroEL vs. GroES), the chain category (cis vs. trans ring), and the residue index, so that residues sharing the same (role, residue index) are assigned to the same group. In heteromeric complexes that contain multiple protein species (e.g. kinase–substrate, small GTPase–effector, receptor–ligand), we explicitly assign distinct role labels to chains belonging to different proteins (e.g. "KRAS" vs. "RAF1", "GroEL" vs. "GroES"). Site-groups are then constructed *only within a fixed role*: residues from different roles never share the same group identifier g_i. Consequently, repeated or symmetry-related copies of the *same* functional site within one protein can be grouped and constrained not to leak across train and test folds, while structurally and functionally unrelated chains in the same PDB entry are evaluated as separate populations.

Formally, each residue is assigned to one of M site-groups:

$$g_i \in \{1, \ldots, M\}, \qquad i = 1, \ldots, N.$$

All residues with the same g_i are treated as a single "functional site" for the purposes of cross-validation.

We then apply a group-aware K-fold splitter (scikit-learn `GroupKFold`) that never places residues from the same site-group into both the training and test folds:

1. We construct fold assignments based on the site-group labels $\{g_i\}_{i=1}^{N}$ using `GroupKFold(n_splits=K)`. By construction, each group $g \in \{1, \ldots, M\}$ is assigned entirely to the training set or entirely to the test set in each fold.

2. For each fold $k \in \{1, \ldots, K\}$:

   (a) Let $\mathcal{I}_{\text{train}}^{(k)}$ and $\mathcal{I}_{\text{test}}^{(k)}$ denote the training and test residue indices returned by `GroupKFold.split`, conditioned on the group labels $\{g_i\}$.

   (b) Fit the mean imputer on $X_{\mathcal{I}_{\text{train}}^{(k)}}$ and transform both training and test feature matrices to obtain $X^{\text{imp}}_{\mathcal{I}_{\text{train}}^{(k)}}$ and $X^{\text{imp}}_{\mathcal{I}_{\text{test}}^{(k)}}$.

   (c) Fit a random forest classifier on $\left(X^{\text{imp}}_{\mathcal{I}_{\text{train}}^{(k)}}, \mathbf{y}_{\mathcal{I}_{\text{train}}^{(k)}}\right)$.

   (d) Obtain predicted class probabilities $\hat{p}_i^{(k)}$ for all $i \in \mathcal{I}_{\text{test}}^{(k)}$ using `predict_proba`.

   (e) Compute the residue-level AUC-ROC on the k-th test fold as before.

3. As above, we summarize performance by the mean and standard deviation of the AUC across folds.

In practice, we adopt the following rule-of-thumb for choosing between residue-level and group-aware splitting. For a given label definition (e.g. T1-all, T1-fullDFT, T1-Ep, or a role-specific subtype), we first construct the site-group labels g_i as described above. If no positive residue shares its $(\text{role}, \text{residue index})$ with another positive residue (i.e. every positive site is unique and does not appear as a replicated copy across chains), then group-aware splitting provides no additional protection against site leakage, and we use `StratifiedKFold` on individual residues. If, however, the positive set includes replicated copies of the same site (multiple residues with identical role and residue index across symmetric chains), we regard these as a single functional site and employ `GroupKFold` to ensure that no such site can appear in both training and test folds. This simple criterion — "use GroupKFold if and only if the label set includes replicated copies of the same $(\text{role}, \text{index})$ site" — guarantees that functional-site leakage is avoided where it matters, while keeping the evaluation protocol simple for tasks without replicated sites.

Optionally, we also report a site-level AUC, in which each site-group g is treated as a single sample. For each test fold k, we aggregate residue-level predictions within site-group g to a single site score $s_g^{(k)}$ (e.g. by taking the mean predicted probability across all residues i with $g_i = g$ in the test set). A site is labeled as functional if any residue in that group has $y_i = 1$, and non-functional otherwise. The AUC-ROC is then computed across $(s_g^{(k)}, y_g)$ pairs. This site-level analysis provides a complementary measure of how well the features distinguish functional vs. non-functional sites, rather than individual residues, under the same no-leakage constraint.

3.2.4 Permutation-based shuffle test for statistical significance

To test whether the observed AUC arises from genuine structure–label association rather than random fluctuations under class imbalance and model flexibility, we perform a permutation (shuffle) test at the residue level with fixed random seeds and a fixed number of permutations.

Hold-out split and observed AUC First, we define a single train–test split consistent with the chosen cross-validation strategy. For simple residue-level evaluation we use a stratified split; for grouped complexes we optionally use a group-aware split so that site-groups are not split across train and test:

1. We construct a train–test split $(X_{\text{train}}, \mathbf{y}_{\text{train}}, X_{\text{test}}, \mathbf{y}_{\text{test}})$ using either `train_test_split` with `stratify = y` (residue-level setting) or a group-aware splitter such as `GroupShuffleSplit` with site-group labels (site-grouped setting). The test fraction is fixed (e.g. `test_size = 0.2`) and the random seed is set to `random_state = 42`.

2. We fit the mean imputer on X_{train} and transform both X_{train} and X_{test} to obtain $X_{\text{train}}^{\text{imp}}$ and $X_{\text{test}}^{\text{imp}}$.

3. We train a random forest classifier with the hyperparameters specified above on $(X_{\text{train}}^{\text{imp}}, \mathbf{y}_{\text{train}})$.

4. We compute predicted probabilities $\hat{p}_i$ for all residues in the test set and the corresponding observed AUC,
$$\text{AUC}_{\text{obs}} = \text{AUC}(\mathbf{y}_{\text{test}}, \hat{p}).$$

Permutation procedure Next, we approximate the null distribution of AUC values that would be obtained if there were no relationship between the features and the labels, while preserving the feature distribution and class proportions. We fix both the number of permutations and the random seed so that the procedure is exactly reproducible.

1. We keep $(X_{\text{train}}^{\text{imp}}, X_{\text{test}}^{\text{imp}}, \mathbf{y}_{\text{test}})$ fixed as obtained above.

2. We set the number of permutations to $B = 120$.

3. We instantiate a NumPy random number generator with a fixed seed, e.g. `rng = numpy.random.RandomState(12345)`, which is used solely for permuting the training labels.

4. For each permutation $b \in \{1, \ldots, B\}$:

   (a) We generate a permuted version of the training labels,
   $$\mathbf{y}_{\text{train}}^{(b)} = \text{permute}(\mathbf{y}_{\text{train}}),$$
   by applying `rng.permutation` to $\mathbf{y}_{\text{train}}$.

   (b) We train a new random forest classifier (with the same hyperparameters and `random_state = 42`) on $(X_{\text{train}}^{\text{imp}}, \mathbf{y}_{\text{train}}^{(b)})$.

(c) We compute predicted probabilities $\hat{p}^{(b)}$ on the unchanged test set and the corresponding null AUC value,

$$\mathrm{AUC}_{\mathrm{null}}^{(b)} = \mathrm{AUC}(\mathbf{y}_{\mathrm{test}}, \hat{p}^{(b)}).$$

5. The collection $\{\mathrm{AUC}_{\mathrm{null}}^{(b)}\}_{b=1}^{B}$ provides an empirical null distribution of AUC values under the hypothesis that labels are independent of the features.

We summarize this null distribution by its empirical mean and standard deviation,

$$\mu_{\mathrm{null}} = \frac{1}{B}\sum_{b=1}^{B} \mathrm{AUC}_{\mathrm{null}}^{(b)}, \qquad \sigma_{\mathrm{null}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\mathrm{AUC}_{\mathrm{null}}^{(b)} - \mu_{\mathrm{null}}\right)^2},$$

and compute a one-sided permutation p-value for the observed AUC as

$$p_{\mathrm{perm}} = \frac{1 + \#\{b : \mathrm{AUC}_{\mathrm{null}}^{(b)} \geq \mathrm{AUC}_{\mathrm{obs}}\}}{1 + B}.$$

With the random seeds, split parameters, and number of permutations fixed as above, the entire evaluation and permutation procedure is exactly reproducible given the same software versions and the same site-group definitions in the grouped setting.

Code availability. The full pipeline for constructing the local environment vectors E_p and the scripts used to reproduce all residue-level analyses in this work are available at `https://github.com/phl043456-star/protein-env-representation` and are permanently archived on Zenodo (`https://doi.org/10.5281/zenodo.17845300`).

4 Results

4.1 Case study: Rubisco large/small subunits (4RUB)

Tier design and functional roles. For the Rubisco holoenzyme (4RUB) we constructed a residue-level Tier 1/Tier 2 scheme on top of the large (L) and small (S) subunits, using only sites with clear experimental or curated functional evidence. Large-subunit chains are denoted A–D and small-subunit chains S–V; all Tier labels are applied per chain by residue index.

Large subunit (L, chains A–D).

- *Tier 1 catalytic core* (`L_Tier1_catalytic_core`): residues that directly participate in RuBP carboxylation or form the immediate catalytic pocket. Per chain this includes

$$\{20, 60, 65, 123, 175, 177, 201, 203, 204, 294, 295, 327, 334, 335\},$$

  expanded across chains A–D for a total of 56 positives.

- *Tier 1 PTM core* (`L_Tier1_PTM_core`): the carbamylated/acetylated lysine K201, treated as the central PTM site (4 positives across A–D).
- *Tier 1 gate/latch* (`L_Tier1_gate_latch`): loop-6 and C-terminal latch positions that gate substrate access and stabilize the closed conformation, represented by R327, K334, and D473 in each chain (12 positives).
- *Tier 1 PPI with Rubisco activase* (`L_Tier1_PPI_Rca_core`): the N-terminal motif (residues 1–4) implicated in Rubisco activase binding. In the processed feature matrix used here, none of these residues are present, so this subset was not evaluated (no positives in `Ep_manifest`).

Tier 2 labels on the L-subunit capture strongly supported but more peripheral roles:

- *Tier 2 catalytic shell* (`L_Tier2_catalytic_shell`): residues forming the immediate structural shell around the catalytic pocket, including 66, 379–381, and 402–404 per chain (28 positives in total).
- *Tier 2 PTM shell* (`L_Tier2_PTM_shell`): residues structurally supporting the PTM core, here K104 and K256 per chain (8 positives).
- *Tier 2 PPI extension* (`L_Tier2_PPI_Rca_extension`): extended N-terminal positions 5–8. As with the Tier 1 PPI core, these residues are not present in the current feature matrix and are therefore not evaluated.

We also define aggregate sets

$$\texttt{L_Tier1_all} := \texttt{L_Tier1_catalytic_core} \cup \texttt{L_Tier1_PTM_core} \cup \texttt{L_Tier1_gate_latch} \cup \texttt{L_Tier1_PPI_}$$
$$\texttt{L_Tier2_all} := \texttt{L_Tier2_catalytic_shell} \cup \texttt{L_Tier2_PTM_shell} \cup \texttt{L_Tier2_PPI_Rca_extension}.$$

as well as a combined catalytic type

$$\texttt{L_Type_catalytic_T1T2} := \texttt{L_Tier1_catalytic_core} \cup \texttt{L_Tier2_catalytic_shell}.$$

and a PTM type

$$\texttt{L_Type_PTM_T1T2} := \texttt{L_Tier1_PTM_core} \cup \texttt{L_Tier2_PTM_shell}.$$

Small subunit (S, chains S–V). On the small subunit, Tier labels emphasize inter-subunit interfaces and helix-8 capping elements:

- *Tier 1 helix-8 cluster* (`S_Tier1_helix8_cluster`): residues in the helix-8 cap that stabilize the holoenzyme and contact the L-subunit, comprising $\{43, 73, 78, 79, 81, 92\}$ per chain (24 positives).
- *Tier 1 βA–βB interface core* (`S_Tier1_interface_betaAB_core`): interface residues S16, L18, and Y32 on each chain that form the βA–βB loop contact surface (12 positives).

Tier 2 labels on the S-subunit capture regulatory and shell positions:

- *Tier 2 βA–βB regulatory* (`S_Tier2_betaAB_regulatory`): positions 59, 67, 68, 69, and 71, which modulate the βA–βB interface (20 positives).
- *Tier 2 helix-8 shell* (`S_Tier2_helix8_shell`): residues 19, 54, and 84 around the helix-8 cluster (12 positives).

We again define aggregate sets

$$\texttt{S_Tier1_all} := \texttt{S_Tier1_helix8_cluster} \cup \texttt{S_Tier1_interface_betaAB_core},$$
$$\texttt{S_Tier2_all} := \texttt{S_Tier2_betaAB_regulatory} \cup \texttt{S_Tier2_helix8_shell}.$$

as well as combined types

$$\texttt{S_Type_betaAB_T1T2} := \texttt{S_Tier1_interface_betaAB_core} \cup \texttt{S_Tier2_betaAB_regulatory},$$
$$\texttt{S_Type_helix8_T1T2} := \texttt{S_Tier1_helix8_cluster} \cup \texttt{S_Tier2_helix8_shell}.$$

Global Tier sets. Finally, we define global Tier sets over the entire complex:

$$\texttt{global_Tier1_all} := \texttt{L_Tier1_all} \cup \texttt{S_Tier1_all}, \qquad \texttt{global_Tier2_all} := \texttt{L_Tier2_all} \cup \texttt{S_Tier2_all}.$$

Performance summary. We evaluated how well the environment representation E_p^{sim} separates each Tier subset from the remaining residues in the 4RUB complex using the random-forest protocol described in Section 2.3.2. Positives are defined by the corresponding Tier set (e.g. `L_Tier1_catalytic_core`); negatives are all other residues present in `Ep_manifest`. For each subset we report both residue-stratified AUC (StratifiedKFold) and site-grouped AUC (GroupKFold with groups defined by the pair role $\in \{\text{L}, \text{S}\}$ and residue index, so that symmetry-related copies of the same site never appear in both train and test folds).

Table 1 summarizes the results. The large-subunit catalytic core and its Tier 2 shell achieve near-perfect separation, even under the stricter site-grouped protocol (AUC ≈ 0.97), indicating that E_p^{sim} captures a highly distinctive environment signature for the entire catalytic basin. In contrast, purely PTM-related shells exhibit weaker site-grouped performance (AUC ≈ 0.44), consistent with the idea that PTM propensity is only partially encoded in the static local environment. On the small subunit, helix-8 and βA–βB clusters are also strongly enriched in the E_p^{sim} feature space (site-grouped AUC $\approx$ 0.72–0.79), although more diffuse regulatory and shell positions are harder to separate. At the global level, functional tiers remain clearly distinguishable from background (AUC ≈ 0.82 for `global_Tier1_all` under GroupKFold), while Tier 2 residues themselves occupy an intermediate, partially separated region of the environment space.

Interpretation. Taken together, the Rubisco case study illustrates a clear hierarchy in how E_p^{sim} resolves functional residues. On the large subunit, the catalytic core and its immediate shell behave as a single, highly distinctive basin in environment space: even when symmetry-related copies of the same active site are forced into different folds by GroupKFold, the classifier retains group-wise AUC values near 0.97 for both the Tier 1 catalytic

Table 1: Residue-level ROC–AUC for Rubisco (4RUB) using E_p^{sim} and a RandomForest classifier. Positives are defined by the indicated Tier subsets; negatives are all remaining residues. Means and standard deviations are computed over k-fold StratifiedKFold (residue-level) and GroupKFold (site-grouped) as described in Section 2.3.2.

Subset	n_+	n_-	$\mathrm{AUC}_{\mathrm{strat}}$	$\mathrm{AUC}_{\mathrm{group}}$
L_Tier1_catalytic_core	56	2308	0.999 ± 0.002	0.973 ± 0.047
L_Tier1_PTM_core	4	2360	1.000 ± 0.000	–
L_Tier1_gate_latch	12	2352	1.000 ± 0.000	–
L_Tier2_catalytic_shell	28	2336	0.999 ± 0.001	0.972 ± 0.045
L_Tier2_PTM_shell	8	2356	0.996 ± 0.005	0.444 ± 0.027
L_Tier1_all	60	2304	0.996 ± 0.006	0.934 ± 0.083
L_Tier2_all	36	2328	0.999 ± 0.001	0.881 ± 0.131
L_Type_catalytic_T1T2	84	2280	0.998 ± 0.002	0.973 ± 0.042
L_Type_PTM_T1T2	12	2352	0.993 ± 0.015	0.447 ± 0.019
S_Tier1_helix8_cluster	24	2340	0.999 ± 0.000	0.789 ± 0.165
S_Tier1_interface_betaAB_core	12	2352	0.999 ± 0.001	0.623 ± 0.291
S_Tier2_betaAB_regulatory	20	2344	0.945 ± 0.033	0.710 ± 0.182
S_Tier2_helix8_shell	12	2352	0.998 ± 0.003	0.459 ± 0.017
S_Tier1_all	36	2328	0.964 ± 0.078	0.783 ± 0.143
S_Tier2_all	32	2332	0.997 ± 0.004	0.550 ± 0.152
S_Type_betaAB_T1T2	32	2332	0.997 ± 0.005	0.719 ± 0.116
S_Type_helix8_T1T2	36	2328	0.965 ± 0.072	0.750 ± 0.170
global_Tier1_all	96	2268	0.984 ± 0.031	0.816 ± 0.109
global_Tier2_all	68	2296	0.982 ± 0.034	0.722 ± 0.173

core and the combined catalytic type `L_Type_catalytic_T1T2`. In contrast, PTM-related shells show almost no separability at the site-grouped level (group-wise AUC ~ 0.44 for `L_Tier2_PTM_shell` and `L_Type_PTM_T1T2`), indicating that the presence or absence of Lys carbamylation/acetylation is only weakly reflected in the static local density/geometry signal captured by E_p^{sim}.

On the small subunit, interface and capping elements occupy an intermediate regime. The combined helix-8 and βA–βB types (`S_Type_helix8_T1T2`, `S_Type_betaAB_T1T2`) achieve group-wise AUC values around 0.7–0.8, showing clear enrichment relative to bulk background, whereas the most diffuse shells (`S_Tier2_helix8_shell`, `S_Tier2_all`) are much harder to separate. At the level of the entire 4RUB complex, both global Tier sets are still resolved above random (group-wise AUC ~ 0.82 for `global_Tier1_all` and ~ 0.72 for `global_Tier2_all`), reinforcing the view that E_p^{sim} is well matched to detecting tight, DFT-like local reactivity pockets and their immediate structural support, while more diffuse regulatory and PTM-specific effects lie closer to the limits of what a static local environment representation can capture.

4.2 Case study: GroEL/GroES chaperonin complex (2C7C)

As a second large, symmetric test system, we applied the residue-level evaluation protocol to the bacterial GroEL/GroES chaperonin complex (PDB ID: 2C7C). The structure consists of a double-ring GroEL tetradecamer that provides an ATP-driven folding chamber and a single-ring GroES heptamer that caps the chamber and contributes to substrate recognition. Across the full complex the environment representation yields a residue feature matrix $X \in \mathbb{R}^{N \times d}$ with $N = 7{,}987$ residues.

Tier-1 labels and grouping. For GroEL/GroES we curated a Tier 1 set that explicitly distinguishes ATPase, substrate-binding, hinge, and inter-ring sites in GroEL (EL) from the interface anchor and extended loop of GroES (ES). Residues are indexed by pairs of a chain identifier and a residue number of the form "chain:resno" (e.g. "A:30"). To define coarse-grained functional roles, we first infer whether each chain belongs to GroEL or GroES by its length: chains whose maximum residue index exceeds 400 are assigned the role label "EL" (GroEL ring), whereas chains with shorter lengths are assigned the role label "ES" (GroES ring). For residue i we denote this role label by $\text{role}_i \in \{\text{EL}, \text{ES}\}$ and its residue number by resno_i.

Within GroEL, Tier 1 is partitioned into four classes. Class A collects ATPase-core residues at the nucleotide-binding pocket that are expected to be highly sensitive to the local electronic environment. In each EL chain, this class includes the N-terminal loop and pocket residues

$$\{30, 31, 32, 33, 34, 51, 52, 87, 89, 398\},$$

corresponding to T30, the 31–33 loop, the conserved loop residue 34, the lysine at position 51, the canonical ATPase residues D52, D87, T89, and the catalytic glutamate at 398. These positions frame the nucleotide-binding site and participate in ATP coordination and hydrolysis; across the 14 EL chains this yields $10 \times 14 = 140$ ATPase-core Tier 1 residues in Class A.

Class B comprises the apical-domain hydrophobic patch that interacts with unfolded substrates and the GroES mobile loop. In each EL chain, we label residues

$$\{199, 201, 203, 204, 234, 237, 263, 264\},$$

for a total of $8 \times 14 = 112$ hydrophobic-patch Tier 1 residues in Class B.

Class C captures the hinge between equatorial and apical domains. The hinge core (Class C_{core}) is represented by residue

$$\{192\}$$

in each EL chain, yielding $1 \times 14 = 14$ hinge-core Tier 1 residues. The hinge-support class (Class $\text{C}_{\text{support}}$) includes residues

$$\{374, 375\}$$

in each EL chain, which stabilize the bending motion, for a total of $2 \times 14 = 28$ hinge-support Tier 1 residues.

Class D covers the inter-ring and inter-subunit contacts along the equatorial interface. In the implementation used for the present evaluation, we define the GroEL inter-ring Tier 1

class by the residue numbers

$$\{105, 197, 386, 434, 452, 461\}$$

in each EL chain. These six positions per chain correspond to the equatorial interface and ring–ring contact region, including residues that form salt bridges and packing interactions between subunits and between the two GroEL rings. Because the same residue-number pattern is applied uniformly to all 14 EL chains that are classified as GroEL, this definition yields $6 \times 14 = 84$ inter-ring Tier 1 residues in Class D. No additional GroEL residues are included in the D-class in the evaluation code; the D-class AUC reported below is therefore computed solely from this set of 84 inter-ring residues versus all remaining EL residues.

Within GroES, Tier 1 is split into two classes. Class E (IVL core) contains the three-residue "IVL" motif (residues

$$\{25, 26, 27\}$$

in each ES chain) that forms the primary anchor to the GroEL apical domain, yielding $3 \times 7 = 21$ IVL-core Tier 1 residues in Class E. Class F (loop support) comprises the remaining mobile-loop residues adjacent to the IVL motif, namely residues

$$\{23, 24, 28, 29, 30, 31\}$$

in each ES chain, which contact substrates and the GroEL cavity but are less tightly constrained than the core IVL anchor. This definition yields $6 \times 7 = 42$ loop-support Tier 1 residues in Class F.

In total, the GroEL Tier 1 set contains 378 residues (140 in Class A, 112 in Class B, 14 in Class $\mathrm{C}_{\mathrm{core}}$, 28 in Class $\mathrm{C}_{\mathrm{support}}$, and 84 in Class D), while the GroES Tier 1 set contains 63 residues (21 in Class E and 42 in Class F), for a global Tier 1 count of 441 residues. All remaining residues in the complex are labeled as background.

As in the Rubisco case, we use a site-grouped splitting protocol to prevent symmetry-related copies of the same functional site from leaking between training and test sets. For each residue i we construct a group label

$$g_i = (\mathrm{role}_i, \mathrm{resno}_i),$$

so that, for example, all GroEL residues with $\mathrm{role}_i = \mathrm{EL}$ and $\mathrm{resno}_i = 398$ share the group label "EL:398" across all EL chains (both rings), and similarly all GroES residues with $\mathrm{role}_i = \mathrm{ES}$ and $\mathrm{resno}_i = 25$ share the group label "ES:25" across all ES chains. Site-groups are constructed independently within EL and ES roles; GroEL and GroES residues never share a group. These group identifiers are passed to `GroupKFold` so that no symmetry-related copy of a functional site appears in both the training and test folds.

Implementation details. The downstream evaluation strictly follows the global protocol in Section 2.3.2. From the feature matrix X and the Tier 1 labels described above, we perform site-grouped cross-validation with `GroupKFold`, using group labels $g_i = (\mathrm{role}_i, \mathrm{resno}_i)$ as defined above. When at least five distinct positive site groups are available for a given label set, we use `GroupKFold(n_splits = 5)`; for smaller subsets we automatically reduce the number

of splits such that each fold contains at least two site groups with both positive and negative residues. In each fold we fit a column-wise mean imputer (`SimpleImputer(strategy = "mean")`) on the training residues and apply it to both training and test residues in that fold. On the imputed features X^{imp} we train a random forest classifier (`RandomForestClassifier`) with

$$\texttt{n_estimators} = 100, \quad \texttt{class_weight} = \text{"balanced"}, \quad \texttt{random_state} = 42, \quad \texttt{n_jobs} = 1.$$

We obtain predicted class probabilities $\hat{p}_i = \hat{\mathbb{P}}(y_i = 1 \mid \mathbf{x}_i)$ on the held-out test residues in each fold, compute the residue-level AUC-ROC per fold, and summarize performance by the mean and standard deviation across folds. For label sets with too few positive site groups to form valid folds (for example, the GroEL hinge-core residue 192 alone), no AUC is reported. The same protocol is applied to (i) the global Tier 1 set (GroEL and GroES combined), (ii) the GroEL-only and GroES-only Tier 1 sets, and (iii) each functional subclass (A, B, C_{core}, $\text{C}_{\text{support}}$, D in GroEL and E, F in GroES).

Global separation and domain-level performance. Using the full GroEL/GroES Tier 1 set as positives and all other residues as negatives, the site-grouped cross-validation yields a global residue-level AUC-ROC of

$$\text{AUC}_{\text{Global, Tier1}} \approx 0.85 \pm 0.05$$

(mean $\pm$ standard deviation across folds). Restricting the analysis to GroEL only, with GroEL Tier 1 residues as positives and all remaining EL residues as negatives, gives

$$\text{AUC}_{\text{GroEL, Tier1}} \approx 0.82 \pm 0.06,$$

while the corresponding analysis on GroES yields

$$\text{AUC}_{\text{GroES, Tier1}} \approx 0.90 \pm 0.11.$$

The somewhat higher AUC in GroES reflects the relatively simple partition between the IVL anchor/loop and the remaining scaffold of the small heptamer, whereas the GroEL Tier 1 class mixes ATPase-site, substrate-patch, hinge, and inter-ring residues that span a wider diversity of structural contexts.

Class-resolved resolution within GroEL and GroES. To quantify which types of functional residues are best captured by the local environment representation, we repeated the analysis for each functional subclass separately, using the corresponding class as positives and all other residues in the same role (EL or ES) as negatives.

Within GroEL, the extended ATPase core (Class A; residues 30, 31, 32, 33, 34, 51, 52, 87, 89, and 398 in each EL chain; 140 residues total) yields an AUC of

$$\text{AUC}_{\text{GroEL, A}} \approx 0.89 \pm 0.08.$$

The apical hydrophobic patch (Class B; residues 199, 201, 203, 204, 234, 237, 263, and 264 in each EL chain; 112 residues total) is separated even more strongly, with

$$\text{AUC}_{\text{GroEL, B}} \approx 0.93 \pm 0.07.$$

For the hinge region, the hinge-core residue set (Class C_{core}; residue 192 in each EL chain; 14 residues total) is too small to provide enough positive site groups for a stable grouped cross-validation, so no AUC is reported for Class C_{core}. In contrast, the hinge-support residues (Class C_{support}; residues 374 and 375 in each EL chain; 28 residues total) yield

$$\text{AUC}_{\text{GroEL, C_support}} \approx 0.93 \pm 0.06.$$

The inter-ring and inter-subunit interface residues (Class D; residues 105, 197, 386, 434, 452, and 461 in each EL chain; 84 residues total) are more challenging: the corresponding AUC is

$$\text{AUC}_{\text{GroEL, D}} \approx 0.79 \pm 0.19,$$

with substantial fold-to-fold variability.

Within GroES, the IVL anchor (Class E; residues 25, 26, and 27 in each ES chain; 21 residues total) is detected with very high accuracy:

$$\text{AUC}_{\text{GroES, E}} \approx 0.95 \pm 0.05.$$

The extended loop (Class F; residues 23, 24, 28, 29, 30, and 31 in each ES chain; 42 residues total) is more heterogeneous in structural role, and its separation performance is correspondingly lower and more variable,

$$\text{AUC}_{\text{GroES, F}} \approx 0.80 \pm 0.19.$$

Interpretation. Taken together, the GroEL/GroES case study confirms that the environment representation is not uniformly sensitive to all kinds of functional importance but exhibits a clear internal hierarchy that tracks the expected dependence on local electronic structure. Within GroEL, residues in the ATPase pocket (Class A), the hydrophobic substrate/GroES-binding patch (Class B) and the hinge-support positions (Class C_{support}) all reach grouped AUC values around 0.9 (typically 0.89–0.93), while in GroES the IVL anchor (Class E) is detected with even higher accuracy (grouped AUC ~ 0.95). These classes occupy a highly distinctive region of the environment space, consistent with their roles as sites of strong local field and density perturbation. In contrast, residues whose roles rely more on distributed mechanical or network properties, such as the GroEL inter-ring contacts (Class D) and the more heterogeneous GroES loop positions (Class F), show only moderate separation with lower grouped AUC values (typically ~ 0.8) and larger variability across folds.

This pattern is consistent with the intended design of the representation as a local, DFT-like surrogate of the electronic and geometric environment: it excels at highlighting residues whose function is tightly tied to strongly perturbed local fields and densities (ATPase chemistry, apical hydrophobic patch, hinge support and IVL anchor), while providing only partial information about residues whose importance emerges primarily from long-range allosteric coupling or global topology. Because the entire protocol (role assignment, group labels g_i, imputation, classifier hyperparameters, cross-validation setup, and label definitions) is fixed as described above, the GroEL/GroES AUC values can be reproduced exactly given the same feature matrices and software versions, and they provide a complementary benchmark to Rubisco for probing the resolution of local environment features in large symmetric complexes.

4.3 Case study: p53–DNA complex (1TSR)

To probe how the environment representation behaves on a multi-functional tumor suppressor with mixed DNA-binding, metal coordination and protein–protein interfaces, we performed a residue-level case study on the p53 core domain bound to its consensus response element (PDB ID: 1TSR). In this structure three copies of the p53 DNA-binding domain form a trimeric assembly around a 21 bp double-stranded DNA, with chains A–C corresponding to the protein domains and chains E–F to the two DNA strands. The environment model again produces a residue feature matrix $X \in \mathbb{R}^{N \times d}$ with $N = 1014$ residues across all protein and DNA chains.

Protein-side Tier-1 labels and grouping. On the protein side, we defined four Tier 1 subsets and a Tier 2 structural shell based on structural, mutational and interaction data for p53. The DNA-contact core Tier 1a comprises residues that form direct base- or backbone-specific contacts in the major or minor groove (Lys120, Arg241, Arg248, Arg273, Ala276, Cys277, Arg280, Arg283). The Zn-coordinating structural cluster Tier 1b collects the metal ligands and their immediate partners (Cys171, His176, Cys179, Cys238, Cys242) that maintain the integrity of the DNA-binding fold. The fitness and stability core Tier 1c contains residues that are not primary DNA ligands but are recurrently depleted in deep mutational scanning data and structural analyses (Arg175, Tyr220, Gly245, Gly249, Glu282). Finally, the allosteric and protein–protein pivot Tier 1d extends the functional set to positions that control cooperativity or form validated binding interfaces with partners such as ASPP2 and BCL-xL: this includes N-terminal surface residues (Gln100, Gln104, Tyr107), L2/L3 loop positions and their allosteric neighbors (His178, Glu180, Arg181, Asp207–Glu213, Met243, Asn247) and a C-terminal side-chain contact with BCL-xL (Asn268). All remaining residues that form a tight structural shell around these cores but have weaker or more indirect functional evidence are collected into Tier 2 (109, 119, 121, 145, 146, 246, 281, 285).

When expanded across chains A–C, these sets yield 102 Tier 1 residues and 126 Tier 1+Tier 2 residues in total. As in the Rubisco case, the Tier 1 labels explicitly include all symmetry-related copies of a given residue number. A naïve residue-level split would therefore leak different copies of the same functional site across training and test sets. To avoid this, we again use a site-grouped protocol. Each residue identifier has the form "chain:resno" (e.g. "B:248"), and for the protein chains A–C we define a group label

$$g_i^{\text{prot}} = (\texttt{P53}, \text{resno}_i),$$

so that, for example, all residues A:248, B:248 and C:248 share the same group label "P53:248". These group labels are passed to `GroupKFold` so that no copy of a functional site appears simultaneously in the training and test folds.

DNA-contact tiers and nucleotide selection. On the DNA side, we introduced an analogous two-tier labeling derived directly from the 1TSR coordinates and the set of key p53 DNA-contact residues. We first identified five central p53 residues that dominate DNA readout in structural and biochemical studies (Lys120, Arg248, Arg273, Arg280, Arg283). For each copy of these residues in chains A–C we enumerated all DNA nucleotides in chains E

and F whose heavy atoms lie within 4.0 Å of any heavy atom of the p53 side chain. Each DNA nucleotide was then assigned a contact multiplicity equal to the number of distinct p53 contact residues among this set.

We define the Tier 1 DNA-contact core as the nucleotides that are contacted by at least two of the five key p53 residues. In 1TSR this procedure yields three nucleotides: E:12 (a thymine contacted by Arg248, Arg273 and Arg280), F:7 (a guanine contacted by Lys120 and Arg283) and F:9 (a guanine contacted by Lys120 and Arg280). These positions sit in the central response-element motif and form a compact cluster where major-groove base readout, minor-groove insertion and backbone recognition coincide. The Tier 2 DNA-contact shell consists of nucleotides that are contacted by exactly one of the five key residues, forming a looser halo around the core. In 1TSR this shell includes E:1, E:11, E:13, E:14 and F:6, F:8, F:14, F:16, which participate in peripheral backbone or groove contacts but do not integrate multiple p53 interaction modes.

For evaluation we treat these DNA positions as nucleotide-level labels: each DNA residue identifier is of the form "chain:resno" (e.g. "E:12"), and Tier 1 or Tier 2 membership is assigned to the corresponding indices in X. In contrast to the protein chains, the DNA chains E–F are not replicated copies, so the grouping is simply at the nucleotide level,

$$g_i^{\mathrm{DNA}} = (\texttt{DNA}, \mathrm{chain}_i, \mathrm{resno}_i).$$

Implementation details and exception handling. The downstream evaluation follows the global protocol of Section 2.3.2 with minor adaptations for small positive sets on the DNA side. From the feature matrix X and the protein and DNA label vectors defined above, we perform site-grouped cross-validation using `GroupKFold`. For the protein tiers we use five folds; for tiers with a smaller number of positive sites the effective number of folds is reduced to the number of distinct positive groups. In each fold we apply a column-wise mean imputer (`SimpleImputer(strategy = "mean")`) fitted on the training residues only and then applied to both training and test residues. On the imputed features X^{imp} we train a random forest classifier

$$\texttt{n_estimators} = 100, \quad \texttt{class_weight} = \text{"balanced"}, \quad \texttt{random_state} = 42, \quad \texttt{n_jobs} = 1,$$

and obtain predicted class probabilities $\hat{p}_i = \hat{\mathbb{P}}(y_i = 1 \mid \mathbf{x}_i)$ on the held-out test residues in each fold.

To ensure that the ROC-AUC is only computed in statistically meaningful settings, we enforce two additional constraints at the fold level. First, the training set for a given fold must contain both positive and negative samples; folds in which the training labels collapse to a single class are skipped to avoid degenerate classifiers that predict a constant probability. Second, the test set must also contain both classes, since ROC-AUC is undefined otherwise; folds violating this condition are likewise skipped. The reported mean and standard deviation are computed over the remaining valid folds. For tiers with extremely few positive sites (notably the three-site DNA Tier 1 core), these criteria can eliminate all folds, in which case the AUC is treated as undefined and not reported.

Global separation across protein tiers. Using the full protein Tier 1 set (Tier 1a ∪ Tier 1b ∪ Tier 1c ∪ Tier 1d; 102 residues) as positives and all other protein residues in

chains A–C as negatives, the site-grouped evaluation yields a residue-level AUC-ROC of

$$\mathrm{AUC}_{\mathrm{p53,Tier1}} \approx 0.89 \pm 0.03.$$

When Tier 2 residues are added to the positive class (126 residues total), the AUC remains essentially unchanged,

$$\mathrm{AUC}_{\mathrm{p53,Tier1+2}} \approx 0.89 \pm 0.02,$$

indicating that the structural shell around the core does not substantially dilute the separation.

Within Tier 1, the different functional subsets show distinct behavior. The DNA-contact core Tier 1a (24 residues) yields

$$\mathrm{AUC}_{\mathrm{p53,Tier1a}} \approx 0.87 \pm 0.07,$$

while the Zn-coordination cluster Tier 1b (15 residues) reaches

$$\mathrm{AUC}_{\mathrm{p53,Tier1b}} \approx 0.90 \pm 0.10,$$

consistent with a strong and localized metal-field signature. The fitness and stability core Tier 1c (15 residues) is more heterogeneous, with

$$\mathrm{AUC}_{\mathrm{p53,Tier1c}} \approx 0.70 \pm 0.14,$$

reflecting the fact that some fitness-critical positions are structurally remote or electronically weak in the local environment representation. In contrast, the allosteric and protein–protein pivot Tier 1d (48 residues) remains well separated,

$$\mathrm{AUC}_{\mathrm{p53,Tier1d}} \approx 0.88 \pm 0.06,$$

indicating that residues controlling loop motion and partner binding often occupy geometrically and electrostatically distinctive environments. Even the Tier 2 structural shell alone (24 residues) exhibits appreciable separation from the background,

$$\mathrm{AUC}_{\mathrm{p53,Tier2}} \approx 0.84 \pm 0.10,$$

suggesting that the immediate neighborhood of the core is also partially resolved in the feature space.

DNA-side separation and contact shell. On the DNA side, the extremely compact Tier 1 DNA-contact core comprises only three nucleotides (E:12, F:7, F:9). Under the conservative grouped-cross-validation protocol and the requirement that both training and test folds contain both classes, no valid fold remains for this set, so we do not report a numeric AUC for the Tier 1 core alone. When we broaden the positive class to the Tier 2 DNA-contact shell (eight nucleotides), treating all other DNA nucleotides in chains E–F as negatives, the nucleotide-level AUC-ROC reaches

$$\mathrm{AUC}_{\mathrm{DNA,Tier2}} \approx 0.91 \pm 0.10.$$

For the union of core and shell nucleotides (11 positions total), the separation further improves to

$$\mathrm{AUC}_{\mathrm{DNA,Tier1+2}} \approx 0.96 \pm 0.05.$$

Despite the small number of positive groups and correspondingly larger fold-to-fold fluctuations, these values indicate that the environment features are sufficient to pull out the DNA segment that sits at the intersection of major-groove base readout, minor-groove insertion and backbone recognition.

Interpretation. Taken together, the 1TSR case study shows that the environment representation can simultaneously resolve multiple layers of functional structure in the p53–DNA complex. On the protein side, it preferentially highlights the Zn^{2+} cluster, the direct DNA-contact core and the most critical allosteric and protein–protein pivots, while still assigning elevated scores to the structural shell around these sites. On the DNA side, a purely geometry- and field-based representation, without using evolutionary profiles or motif-level supervision, is able to distinguish the narrow band of nucleotides engaged by the key p53 contact residues from the remainder of the duplex. The explicit handling of symmetry via site-grouped cross-validation and the conservative fold-level filtering for degenerate splits ensure that the reported AUC values reflect genuine separation in feature space rather than artefacts of overfitting, label leakage or unstable folds in tiers with very few positive sites.

4.4 case study: SecA ATPase motor and preprotein clamp (1TF5)

Label design. For SecA (1TF5, chain A) we constructed a three-way partition of functional residues into *ATPase*, *clamp* and *C-terminal relay/interface* modules, each with a strict Tier 1 "core" and a Tier 2 "support" shell. All labels are defined in 1TF5 PDB numbering and applied only to chain A; the remaining chains in the environment matrix are treated as unlabeled background.

The **ATPase module** is anchored on the canonical Walker A/B motifs and the nearby catalytic residues in the N-terminal domain. The Tier 1 core contains six residues (106, 107, 207, 208, 215, 216) that directly support ATP binding and hydrolysis: the P-loop Lys and its immediate neighbour, the Walker B acidic pair, and a downstream acidic cluster that coordinates the catalytic water and metal ion. Around this core we define a Tier 2 ATPase shell consisting of the flanking loops and helices (100–105, 108–119, 200–206, 209–214, 217–220, minus the core positions), capturing the structural context of the nucleotide pocket without conflating it with the strict catalytic sites.

The **clamp module** corresponds to the preprotein-binding groove in the PPXD/helical scaffold region. Here the Tier 1 core is a small set of hydrophobic residues on two helices (357, 360, 361, 381, 384, 385) that directly line the groove and form the main binding surface for signal peptides and unfolded segments. The Tier 2 clamp shell extends this to the full two-helix bundle and its adjacent loops (352–356, 357–362, 375–380, 377–386, 387–389, again excluding the core residues), so that we can separately test whether the environment vector E_p prioritizes the central grip versus its immediate geometric neighbourhood.

Finally, the **C-terminal relay/interface module** covers the extended C-terminal helix bundle that participates in SecA–SecYEG and SecA–SecA interactions and includes the base

of the two-helix finger. The Tier 1 core here is defined as a set of seven aromatic and polar anchors (590, 599, 611, 615, 642, 652, 665) distributed along the interface helices and finger loop. The Tier 2 interface shell spans the surrounding helices and loops (571–619, 623–642, 643–655, 656–663, 664–674, excluding the core sites), which together form the relay scaffold transmitting conformational changes from the ATPase motor towards the membrane channel.

From these module-wise definitions we obtain three aggregate label collections: (i) the union of all Tier 1 cores (19 residues) as a global functional core set, (ii) the union of all Tier 2 shells (149 residues) as a broad support/relay set, and (iii) the combined Tier 1+Tier 2 union (168 residues) as a composite functional surface. All evaluations use $X := E_p^{\text{sim}}$ (10 geometric plus physico-chemical features), a RandomForest classifier with class_weight=balanced, and GroupKFold with site-groups $g_i = \text{SecA} : \text{resseq}_i$ to avoid leakage between symmetry-related copies of the same residue.

Residue-level performance. With the pooled Tier 1 core union as the positive class (19 positives vs 885 negatives), the environment vector E_p yields a residue-level ROC–AUC of

$$\text{AUC}_{\text{Tier1 core, all}} = 0.97 \pm 0.03$$

(mean ± s.d. over 5 GroupKFold splits). When we resolve this core by functional module, the ATPase, clamp and C-terminal relay all remain strongly separable. For the *ATPase core* (6 positives; Walker A/B and catalytic cluster) we obtain

$$\text{AUC}_{\text{ATPase, core}} = 0.86 \pm 0.22,$$

indicating that residues directly responsible for ATP binding and hydrolysis sit near the top of the E_p score distribution. The *clamp core* (6 positives) shows a very similar behaviour, with

$$\text{AUC}_{\text{clamp, core}} = 0.86 \pm 0.23,$$

consistent with a clear physical signal for the hydrophobic grip that engages preproteins. The *C-terminal relay/interface core* (7 positives) is even more sharply distinguished:

$$\text{AUC}_{\text{interface, core}} = 0.98 \pm 0.01,$$

showing that the key aromatic and polar anchors in the C-terminal helix bundle are almost perfectly separated from the global background by a purely local environment descriptor.

Moving from strict cores to the Tier 2 shells, the model continues to strongly enrich for the functional regions while naturally assigning slightly lower scores than for the Tier 1 sites. For the union of all Tier 2 support residues (149 positives vs 755 negatives) we obtain

$$\text{AUC}_{\text{Tier2 support, all}} = 0.94 \pm 0.02,$$

with module-wise values

$$\text{AUC}_{\text{ATPase, shell}} = 0.94 \pm 0.04, \quad \text{AUC}_{\text{clamp, shell}} = 0.95 \pm 0.07, \quad \text{AUC}_{\text{interface, shell}} = 0.95 \pm 0.02.$$

Thus residues that support the ATPase pocket, clamp helices, and C-terminal relay—but are not themselves strict hotspots—are still consistently ranked above the bulk surface.

When we consider the full Tier 1+Tier 2 union (168 positives vs 736 negatives), the separation remains high:

$$\mathrm{AUC}_{\text{Tier1+2, all}} = 0.95 \pm 0.02,$$

and the same trend is seen when aggregating cores and shells within each module:

$$\mathrm{AUC}_{\text{ATPase, core+shell}} = 0.96{\pm}0.02, \quad \mathrm{AUC}_{\text{clamp, core+shell}} = 0.95{\pm}0.07, \quad \mathrm{AUC}_{\text{interface, core+shell}} = 0.95{\pm}0.02.$$

In other words, as we enlarge the positive class from the minimal hotspots to the entire functional module, the environment vector continues to track the biologically important surface without collapsing towards random performance.

Shuffle control. To exclude the possibility that these high AUC values arise from label–feature leakage or from artefacts in the cross-validation procedure, we performed a label-shuffling control. Here the residue labels are randomly permuted while keeping E_p^{sim} and the site-group assignments fixed, and the same GroupKFold evaluation is repeated over multiple random permutations. For the Tier 1 core union the shuffled labels give

$$\mathrm{AUC}^{\text{shuffle}}_{\text{Tier1 core}} = 0.47 \pm 0.07,$$

while the Tier 1+2 union and the Tier 2 support-only set yield

$$\mathrm{AUC}^{\text{shuffle}}_{\text{Tier1+2}} = 0.50 \pm 0.04, \qquad \mathrm{AUC}^{\text{shuffle}}_{\text{Tier2 support}} = 0.50 \pm 0.03,$$

all statistically indistinguishable from random ranking. This control confirms that the SecA performance is not an artefact of the evaluation pipeline: a purely local environment vector E_p is sufficient to recover, with high specificity, the ATPase motor, the preprotein-binding clamp and the C-terminal relay/interface module in a large, asymmetric ATPase motor dimer.

Interpretation. Taken together, the SecA case study shows that the environment representation exhibits a clear internal hierarchy rather than a uniform sensitivity to "importance". The pooled Tier 1 core set (ATPase pocket, clamp hydrophobics, and C-terminal relay/interface anchors) is separated from the global background with AUC ≈ 0.97, indicating that these residues occupy an extremely distinctive region of environment space. At the module level, the ATPase core and clamp cores already reach high AUC values (~ 0.86), while the C-terminal relay/interface core is almost perfectly separated (~ 0.98). When we move from these strict cores to the Tier 2 support shells, the separation remains high across all modules (AUC $\sim$ 0.94–0.95), and the combined Tier 1+Tier 2 sets preserve this performance (~ 0.95 overall and per class). Thus residues that are biochemically identified as catalytic or relay-critical, as well as those that merely buttress the ATPase, clamp, or interface modules, all cluster in a compact and well-separated region of the local environment descriptor.

This pattern is consistent with the intended design of the representation as a local, DFT-like surrogate of the electronic and geometric environment: it strongly highlights residues whose role is tightly coupled to perturbed local fields and densities (ATPase chemistry,

preprotein clamp, C-terminal relay), and it continues to distinguish their immediate structural neighbourhood, while saying nothing special about the shuffled controls (which yield AUC ≈ 0.5). The main caveat is that our Tier 1/Tier 2 sets are structure- and literature-based supersets of well-characterised residues rather than an exhaustive catalogue; some bias toward "obvious" sites is therefore unavoidable. Within that limitation, SecA provides a stringent but favourable test: without using any evolutionary information or sequence motifs, a purely local environment vector E_p already recovers most of the residue-level functional enrichment in this large ATPase motor, suggesting that remaining failures are more plausibly due to genuinely non-local or context-dependent effects than to missing signal in the local physics encoded by E_p.

5 Discussion

5.1 Where the representation is strong

Across the panel of proteins and complexes examined here, the environment representation E_p consistently shows its strongest performance on residues whose roles are dominated by local electronic structure and short-range geometry. Canonical examples include metal-coordination sites, catalytic cores, tight protein–protein interfaces, and hinge-like mechanical pivots.

In the Rubisco case study, the large-subunit catalytic core and its immediate shell (Tier 1/2 catalytic sets) achieve site-grouped AUROC values in the high 0.8–0.9 range, occasionally approaching near-perfect separation for the most tightly defined cores. These residues sit in well-defined density pockets around the carbamylated Lys and associated acidic cluster, with clear contributions from the pseudo-density, curvature, and metal/phosphate field components of E_p. A similar pattern is observed in the small-subunit helix-8 cap and β-strand interface, where the environment vector separates structurally central interface residues from the surrounding scaffold.

For GroEL/GroES, the environment space sharply distinguishes the ATPase pocket and substrate/GroES-binding hydrophobic patches, as well as a set of hinge-support positions that transmit conformational changes across domains. Again, these classes are characterised by distinctive local density and curvature signatures, reinforced by anisotropic neighbour geometry. The SecA ATPase core and preprotein clamp show analogous behaviour: the Walker-motif catalytic residues and hydrophobic clamp groove reside in a compact, well-separated region of environment space, whereas broader interface regions are more diffuse.

The p53 case study further reinforces the locality of the signal. Residues that make direct DNA contacts, form the Zn^{2+}-binding structural cluster, or lie in the immediate stability core all fall into environment regions that are relatively easy to separate from the global background. In each case, the combination of pseudo-density, external metal field, and local geometric descriptors is sufficient to recover most of the physically obvious "quantum-ish" structure around these sites. Taken together, these results demonstrate that E_p is well aligned with roles that can be described as local electronic or steric motifs embedded in a relatively rigid neighbourhood.

5.2 Where the representation is weak

By contrast, the environment representation performs much less reliably on labels whose definition is inherently global or context-dependent. This includes residues whose functional importance emerges only at the level of fitness landscapes, extended allosteric networks, diffuse scaffolding roles, or highly delocalised mechanical couplings.

In p53, many residues identified by deep mutational scanning or protein–protein interaction studies as functionally critical do not correspond to clean local extrema in E_p. Their roles depend on long-range coupling, multi-partner interfaces, or conformational selection, and the environment vector, by construction, only encodes a single static conformation and its immediate neighbourhood. Similarly, in SecA and GroEL/GroES, extended interface bands and mechanically important shells around the core sites show much weaker and noisier separation. Even when the AUROC remains above random, the variance across folds is high and there is no clear low-dimensional cluster corresponding to these diffuse roles.

More generally, any label that implicitly mixes multiple physical regimes—for example global stability, context-specific regulation, or organism-specific fitness effects—tends to be poorly aligned with the purely local, static, density-based description in E_p. These cases mark the boundary where additional sources of information (dynamics, evolution, or explicit multi-state modelling) become indispensable.

5.3 What the results mean from a representation-only perspective

Viewed purely as a representation, these findings suggest a useful decomposition of functional variance into a "local quantum-informed" component and a residual component that must be explained by other modalities. The environment vector captures those aspects of residue importance that can be inferred from a coarse-grained surrogate for electron density, local curvature, solvation exposure, and short-range packing. When E_p alone achieves AUROC values in the 0.8–0.9 range for a particular Tier, it is reasonable to interpret that Tier as being largely determined by local physics in the chosen conformation.

Conversely, when a Tier exhibits only modest or unstable separation under the same protocol, this is evidence that its defining features lie outside the representational scope of E_p: long-range mechanical networks, multi-state allostery, evolutionary constraints, or cell-level context. In this sense, the panel of case studies functions not only as a performance benchmark but also as an empirical probe of what fraction of a given functional annotation can be attributed to local electronic and geometric structure.

Importantly, the probe classifier is not the primary object of interest. Random Forests are used here as diagnostic tools to measure how linearly and nonlinearly separable different functional subsets are in the environment space. The variation in site-grouped AUROC across Tiers and proteins therefore reflects differences in representation alignment, rather than being a statement about the ultimate ceiling of predictive performance achievable by more sophisticated models.

5.4 Practical implications for structural bioinformatics pipelines

From a practical standpoint, the environment representation is best viewed as a deterministic, reusable pre-processor that can be attached to a wide range of structural learning tasks. Because E_p is computed directly from coordinates and element types, with explicit handling of metal ions and phosphates, it can be precomputed once per structure and supplied as input features to downstream models without retraining the representation itself.

In tasks that are dominated by local physics—metal-binding prediction, catalytic site ranking, core interface hotspot detection, or hinge identification—the present results suggest that simply appending E_p to existing geometric or graph-based encodings should provide immediate gains. In more global tasks, such as allosteric pathway discovery or fitness landscape modelling, E_p can serve as one view among many, complementing evolutionary, dynamical, or sequence-based representations.

Because the pipeline is deterministic and metal-aware, it is also well suited as a standardised front-end for benchmarking. Different machine-learning models can be compared on the same fixed environment vectors, removing ambiguity about how much of their performance comes from representation learning versus downstream classifier capacity. In generative settings, E_p could be used as a physically informed constraint or diagnostic, flagging designs whose local environments are inconsistent with known catalytic or binding motifs.

5.5 Limitations and future directions

Several limitations of the present work are worth emphasising. First, despite the terminology, the representation does not use true *ab initio* DFT densities. The density components in E_p are based on promolecular or otherwise coarse-grained surrogates, chosen for computational tractability and determinism rather than chemical accuracy (as defined in Section 2.2). The results therefore speak to the usefulness of a "DFT-inspired" stencil of local density- and field-like quantities, not to the sufficiency of such surrogates as faithful substitutes for quantum calculations.

Second, the current implementation is entirely static. Dynamics, multiple conformational states and time-dependent fields are not modelled. All case studies are based on single (or a small number of) PDB structures, and E_p reflects only the local environment in those snapshots. For phenomena where functional roles are encoded in rare conformations or in the distribution over states rather than any single structure, the present representation is intrinsically blind.

This limitation points to a natural avenue for future work: unifying local reactivity metrics with global dynamical descriptions of conformational landscapes. Recent work on Riemannian formulations of protein energy landscapes suggests that geodesic structure can capture slow collective motions at a global scale. A future extension of our framework could combine such global geodesic constraints with the local descriptor E_p (e.g. by embedding it in a space of trajectories in $\mathbb{R}^3 \times T$), so that both local fields and large-scale motions contribute to the same notion of reactivity. Exploring this direction lies outside the scope of the present study.

Third, the empirical panel, while diverse, remains limited: a small set of enzymes, chaperonins, transcription factors and transport ATPases. The qualitative patterns we

observe—strong performance on local physicochemical cores and weaker alignment on global or context-dependent labels—are robust across these systems, but they should be retested on larger curated benchmarks before being treated as general. Moreover, our label design is intentionally biased towards mechanisms with clear local structural anchors (catalytic basins, metal clusters, interface anchors, hinge supports). Residues whose importance is dominated by long-range electrostatic steering, intrinsically disordered regions, phase-separating low-complexity segments or global mechanical properties are treated as background in all tasks and thus lie outside the scope of our current resolution analysis.

Finally, the environment space itself can be extended in several directions. A more faithful treatment of electron density, even at semi-empirical or fragment-based levels, could sharpen sensitivity to subtle catalytic or redox effects. Incorporating explicit multi-state information or coarse dynamical descriptors would help bridge the current gap on allosteric and fitness-defined residues. On the mathematical side, the density-overlap metric and associated measure structure open the door to applying tools from optimal transport and topological data analysis directly on the environment manifold. The present work should therefore be viewed as a coarse-grained step towards a family of quantum-informed, geometrically structured representations of macromolecular reactivity, rather than as a complete solution.

A Formal Definition of the Artificial Manifold Class V'

A.1 Overview

This appendix formalizes the construction of an artificial protein manifold class, denoted V'. The class V' provides an idealized, simulation-friendly setting for testing the hypothesis of structure–function separability in a controlled, fully specified environment. In particular, V' allows us, in principle, to generate "ground truth" functional labels directly from a prescribed scoring rule derived from our theory. By knowing exactly how function is encoded in V', one can evaluate whether a given computational pipeline is capable of recovering that known pattern without any experimental noise or annotation ambiguity.

During model development we used variants of V' as an internal sanity check on the feature-computation and learning pipeline, but no quantitative results on V' are reported or relied upon in the main text. The construction below is provided to make this artificial class explicit and self-contained for future work.

A.2 Sampling a Geometric Structure from V'

We define an element of V' as a set of n points ("residues") embedded in $\mathbb{R}^3$, such that each point has a well-defined neighbourhood. In a benchmark configuration we chose $n = 60$; other instances of V' may use different values of n provided they satisfy the rules below.

Let $\{p_1, \ldots, p_n\} \subset \mathbb{R}^3$ denote the points. We endow this discrete set with a topology by declaring that each point p has a neighbourhood

$$U_p := \{\, q \in \{p_1, \ldots, p_n\} : \|q - p\| \leq 2.5 \text{ Å} \,\}.$$

The collection $\{U_p\}$ plays the role of an atlas for V', which we treat as a 2D manifold in the loose sense that each point has a small, locally planar neighbourhood, even though V' is not a smooth continuum. This construction is intended to mirror a protein's Cα trace, where each residue has a local cluster of neighbours within a fixed cutoff.

For convenience we also define, for each residue p, a fragment F_p consisting of p and its neighbours,

$$F_p := \{p\} \cup U_p,$$

which plays the role of the residue-centred fragment used in the full 1–9 environment pipeline.

A.3 Environment Vectors in V'

Each residue p in V' is assigned a theoretical environment vector E_p^{theory} that mirrors the structure of the implemented 1–9 environment pipeline used for real proteins. In the artificial manifold we do not run quantum calculations; instead, we instantiate only the B-mode (surrogate) components, but keep the same layout of environment terms: charge, dipole magnitude, curvature-like density smoothness, element-resolved external-field block, solvent exposure, and a 4D physicochemical embedding.

Formally, the environment vector has the form

$$E_p^{\text{theory}} \;=\; \big(q_p,\ \|d_p\|,\ C_p^{\rho},\ \text{external_field}_p,\ \text{SASA}(p),\ \text{Emb}(p)\big),$$

where $\text{external_field}_p \in \mathbb{R}^3$ and $\text{Emb}(p) \in \mathbb{R}^4$. The individual components are defined below by transplanting the real-case construction into the simplified geometry of V'.

Surrogate density on F_p. We treat each residue p as a pseudo-atom located at its coordinate in $\mathbb{R}^3$. To emulate the B-mode promolecular density used in the real pipeline, we assign to each residue q an element-like label (e.g. "protein-like", "metal-like", "P-like", "halogen-like", "hetero-like") and an associated weight w_q and kernel width σ_q on the scale of atomic radii.

For each residue p, we construct a fragment-centred Cartesian grid over the bounding box of F_p with padding and spacing chosen analogously to the real pipeline (e.g. padding $P \approx 2.0$ Å, grid spacing $\Delta \approx 0.5$ Å, cell volume $\Delta V = \Delta^3$). The surrogate density ρ_p is then defined on this grid by

$$\rho_p(x) \;=\; \sum_{q \in F_p} w_q \exp\!\left(-\frac{\|x - r_q\|^2}{2\sigma_q^2}\right),$$

where r_q is the coordinate of residue q. This is a toy analogue of the B-mode promolecular density on a fragment, but constructed entirely from the point set of V'.

Local integration mask U_p. To approximate the theoretical neighbourhood U_p used in the main model, we define a mask on the fragment grid. Let R_{env} be an environment radius (e.g. 2.0 Å). We set

$$U_p^{\text{grid}} := \big\{x_i : \exists\, q \in \{p\} \text{ s.t. } \|x_i - r_q\| \le R_{\text{env}}\big\},$$

i.e. the grid points lying within R_{env} of any atom in residue p (here, its single pseudo-atom). All integrals below are approximated by Riemann sums over U_p^{grid}.

Density-based descriptors D_p, d_p^ρ, C_p^ρ. Using ρ_p on the fragment grid, we replicate the density-based descriptors from the real pipeline:

- Total local density:

$$D_p \approx \sum_{x_i \in U_p^{\text{grid}}} \rho_p(x_i)\, \Delta V.$$

- Density dipole-like moment relative to the residue centre r_p:

$$d_p^\rho \approx \sum_{x_i \in U_p^{\text{grid}}} \rho_p(x_i)\, (x_i - r_p)\, \Delta V.$$

- Curvature-like smoothness / roughness measure:

$$C_p^\rho \approx \sum_{x_i \in U_p^{\text{grid}}} \|\nabla \rho_p(x_i)\|^2\, \Delta V,$$

  where $\nabla \rho_p$ is evaluated by finite differences on the fragment grid under implicit Neumann boundary conditions. As in the real pipeline, C_p^ρ is interpreted as a smoothness proxy of the surrogate density rather than a geometric curvature of a fixed isosurface.

In the absence of quantum charges, the dipole component in the environment vector uses $\|d_p^\rho\|$ directly:

$$\|d_p\| := \|d_p^\rho\|.$$

Surrogate charge component q_p. The theoretical model uses a residue-level average of the charge field $q(x)$ over U_p. In the artificial manifold we do not run xTB and instead emulate the B-mode residue-level charge surrogate.

We assign to each residue q an element-like label and an element-dependent weight $w_{\text{elem}}(\text{type}(q))$ (negative for electronegative "O-like" or "halogen-like" types, positive for "metal-like" types, zero for neutral background), mirroring the element-based table used in the real pipeline. The base charge for residue p is then

$$Q_p^{\text{atom}} = \sum_{q \in p} w_{\text{elem}}\big(\text{type}(q)\big),$$

which in this toy setting reduces to a single weight attached to the pseudo-atom at p.

To mimic SASA-modulation of charge exposure, we define a simple solvent-exposure proxy from the discrete neighbourhood size:

$$\text{SASA}(p) \propto \frac{1}{|U_p|},$$

where $|U_p|$ is the number of neighbours within the 2.5 Å cutoff, and then normalise $\text{SASA}(p)$ to lie in $[0, 1]$ across the manifold. We obtain a scaling factor

$$s_p = \phi\left(\frac{\text{SASA}(p) - \overline{\text{SASA}}}{\overline{\text{SASA}}}\right),$$

where $\overline{\mathrm{SASA}}$ is the mean over all residues and ϕ is a clipped linear map as in the real B-mode construction. The surrogate residue charge is then

$$Q_p^{\mathrm{B}} = s_p\, Q_p^{\mathrm{atom}},$$

and we set

$$q_p := Q_p^{\mathrm{B}}$$

as the charge component in E_p^{theory}.

Element-resolved external-field block external_field$_p$**.** To remain consistent with the main environment pipeline, we include a compact external-field block that captures the contribution of different "channels" (metal-like, phosphorus-like, halogen-like) to the local density around residue p.

In the real implementation this is done via element-resolved density channels. We reproduce the same construction here at the level of pseudo-atoms. The residue set of V' is partitioned into disjoint subsets

$$\mathcal{V}^{(\mathrm{prot})}, \quad \mathcal{V}^{(\mathrm{metal})}, \quad \mathcal{V}^{(P)}, \quad \mathcal{V}^{(\mathrm{hal})}, \quad \mathcal{V}^{(\mathrm{het})},$$

according to their assigned type labels. For each channel $c \in \{\mathrm{metal}, P, \mathrm{hal}\}$ we define a channel-restricted density on the fragment grid by summing only over residues in $\mathcal{V}^{(c)} \cap F_p$, and compute the channel-resolved local density

$$D_p^{(c)} \;\approx\; \sum_{x_i \in U_p^{\mathrm{grid}}} \rho_p^{(c)}(x_i)\, \Delta V.$$

Given the total local density D_p, we set the fractional contribution

$$f_p^{(c)} = \frac{D_p^{(c)}}{D_p} \quad \text{if } D_p > 0,$$

and $f_p^{(c)} = 0$ otherwise. The external-field block is then

$$\text{external_field}_p := \left\| F_p^{(\mathrm{metal})},\ F_p^{(P)},\ F_p^{(\mathrm{hal})} \right\| := \left\| f_p^{(\mathrm{metal})},\ f_p^{(P)},\ f_p^{(\mathrm{hal})} \right\|.$$

This channel-based variant matches the default choice used in the real pipeline: it depends only on local density composition and does not require any global loop over the entire structure.

SASA component. The scalar SASA(p), defined above from the inverse neighbour count and normalised to $[0, 1]$, is inserted directly as the solvent-exposure component of the environment vector. Thus, in V' it plays the same structural role as the Shrake–Rupley SASA in the real implementation, but is computed from the discrete geometry instead of an explicit molecular surface.

Physicochemical 4D embedding $\mathrm{Emb}(p)$**.** To preserve the layout of the real environment vector, we include the same 4D physicochemical embedding used for real amino-acid residues:

$$\mathrm{Emb}(p) = [q_{\text{formal}},\, h_{\text{hydropathy}},\, s_{\text{steric}},\, a_{\text{aromatic}}] \in \mathbb{R}^4.$$

For V', there is no true sequence, so we assign each residue an artificial type index and map it to a 4D vector via a lookup table that mimics the real residue embedding (e.g. by sampling from the set of actual amino-acid embeddings or from a compressed class-wise version). This ensures that the environment vector in V' has exactly the same component structure and numerical ranges as in the real protein pipeline, while remaining fully controlled and synthetic.

In summary, the artificial manifold V' is equipped with a residue-centred environment vector E_p^{theory} that reuses the same conceptual components as the real 1–9 environment pipeline—charge, dipole magnitude, curvature-like density smoothness, element-resolved external field, solvent exposure, and a 4D embedding—instantiated via surrogate densities and synthetic residue types on a random geometric point cloud.

A.4 Reactive Fiber Assignment in V'

We now specify how "reactive" residues are defined in V'. For a given instance of V', we choose a small cluster of points—for example, 5 out of 60 residues—to be reactive, arranged to be spatially close in order to emulate an active-site region.

For each reactive residue p we attach a non-trivial fibre F_p representing a localised quantum-like state. Concretely, we assign to every residue p a numeric *coherence score* C_p defined by an exponential function centred on the chosen reactive cluster:

$$C_p = \exp\bigl(-\alpha\, d(p, \mathcal{C})^2\bigr),$$

where $\mathcal{C}$ is the set of reactive-centre points, $d(p, \mathcal{C})$ is the distance from p to the cluster (for example, the minimum Euclidean distance to any point in $\mathcal{C}$), and $\alpha > 0$ is a decay parameter. Scores are therefore highest for residues in or near the cluster and decay smoothly away from it. We then declare residues with C_p above a chosen threshold (for example, the 95th percentile of the $\{C_p\}$ values) to be reactive. By construction, we thus obtain "ground truth" reactive vs. non-reactive labels together with a closed-form scoring function that generates them.

For non-reactive residues in V', the associated fibre F_p is taken to be trivial (identity map, no additional data). For reactive residues, F_p can be regarded as storing the coherence score C_p or, more abstractly, an artificial wavefunction. In practice, for the purposes of algorithmic validation, it suffices to know that these residues carry an extra scalar degree of freedom that is perfectly correlated with their reactive status. This makes V' a convenient toy setting in which the relationship between geometry, environment vectors, and functional labels is fully specified.

B Mathematical foundations of the environment–vector space

B.1 Definition as an embedded sub-manifold

Definition B.1 (Environment–vector manifold)**.** Let $(M \times T, g)$ be the 4-D Riemannian manifold of atomic configurations (Section 2.3). The map

$$\iota : M \times T \hookrightarrow \mathbb{R}^k, \quad p \mapsto E_p := \Big[\int \rho_p(x)\psi_m(x)\, d\mathrm{vol}_g(x) \Big]_{m=1}^{k}$$

is a C^1 immersion; its image $E := \iota(M \times T)$, equipped with the pull-back metric $g_E := \iota_*(g)$, is called the *environment–vector manifold*.

B.2 Metric and norm equivalence

Theorem B.2 (Ep metric equivalence)**.** *For all $p, q \in E$,*

$$c\, \|E_p - E_q\|_2 \;\le\; \|\rho_p - \rho_q\|_{L^2} \;\le\; C\, \|E_p - E_q\|_2, \qquad c, C > 0,$$

and $d_\rho(p,q) := \arccos\big[\langle \rho_p, \rho_q \rangle / (\|\rho_p\| \|\rho_q\|)\big]$ induces the same topology as $\|E_p - E_q\|_2$. Consequently (E, d) is complete (Lemma 3.4–Theorem 3.5 in supplement file v2).

B.3 Sub-Gaussian probability measure on E

Lemma B.3 (Tail bound)**.** *With $\mu(dE) = Z^{-1} \exp[-\alpha \|E\|_2^2]\, d\mathrm{vol}_{g_E}$, $\alpha > 0$, we have for all $r > 0$*

$$\mu\big(\|E\|_2 > r\big) \;\le\; \exp\big(-\alpha r^2\big),$$

so every polynomially bounded functional on (E, μ) is integrable (Lemma 4.3 in supplement file v2).

Dimension and normalisation. Throughout, $\dim(M \times T) = r$, so $E = \iota(M \times T)$ is an r-dimensional sub-manifold of $\mathbb{R}^k$. In Lemma E.3 the normalisation constant is $Z = (\alpha/\pi)^{k/2}$, independent of the specific residue.

Motivation. Metric equivalence (Theorem B.2) ensures that switching from the Euclidean kernel used in Eq. (2) to a geodesic kernel (Appendix C) leaves the topology and completeness of (E, d) unchanged.

C Time-Coherent Reactivity

Time-resolved extension of the environment vector

In the main text we define a static environment vector $\mathrm{Ep}_p \in \mathbb{R}^k$ for each residue p, computed from a single protein structure (one PDB/mmCIF model). The components of Ep_p aggregate

local geometric quantities (for example neighbour density, asymmetry, curvature), solvent exposure, and optional sequence- or metal-aware terms within a fixed cutoff around residue p.

In many realistic settings, however, a protein is not represented by a single structure but by a trajectory or ensemble

$$\{X(t_\ell)\}_{\ell=1}^{L},$$

where each frame $X(t_\ell)$ is a full set of atomic coordinates at a time-like index t_ℓ (for example, molecular dynamics, Monte Carlo sampling, or a collection of experimental conformers). In this case we simply apply the same environment-vector construction frame by frame.

Frame-wise environment vectors. For each residue p and each frame t_ℓ we define a neighbourhood $U_p(t_\ell)$ (for example, all atoms within a fixed cutoff radius) and compute the same local descriptors as in the static case: density-based terms, charge integrals, neighbour asymmetry, curvature density, solvent-accessible surface area, and any additional embedding components. This yields a time-dependent environment vector

$$\mathrm{Ep}_p(t_\ell) \in \mathbb{R}^k, \qquad \ell = 1, \ldots, L,$$

with the same dimensionality k and semantic interpretation as the original static Ep_p.

Instead of a single point in environment space, each residue p now traces out a discrete trajectory

$$\Gamma_p \;=\; \big\{\mathrm{Ep}_p(t_\ell)\big\}_{\ell=1}^{L} \;\subset\; \mathbb{R}^k.$$

Static summaries of dynamic environments. To obtain a single descriptor per residue that incorporates all frames, we define simple time-coherent summaries of the Ep trajectory.

- **Time-averaged environment vector**

$$\overline{\mathrm{Ep}}_p := \frac{1}{L}\sum_{\ell=1}^{L} \mathrm{Ep}_p(t_\ell) \in \mathbb{R}^k.$$

  This has the same units and interpretation as the static environment vector and reduces exactly to Ep_p when $L = 1$.

- **Fluctuation magnitude in Ep space**

$$\mathrm{Var}[\mathrm{Ep}_p] := \frac{1}{L}\sum_{\ell=1}^{L} \big\|\mathrm{Ep}_p(t_\ell) - \overline{\mathrm{Ep}}_p\big\|^2,$$

  where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^k$. This scalar quantifies how strongly the local environment of residue p fluctuates over the trajectory.

- **Path length in Ep space**

$$L_p := \sum_{\ell=1}^{L-1} \big\|\mathrm{Ep}_p(t_{\ell+1}) - \mathrm{Ep}_p(t_\ell)\big\|,$$

which is a discrete measure of how far the environment of p moves in Ep space across the trajectory.

These dynamic quantities can be combined with the static scoring terms defined elsewhere in the manuscript to obtain a time-coherent residue-level reactivity or functional score. For example, one may up-weight residues whose static environment is strongly reactive and whose environment vector exhibits large, structured fluctuations over the trajectory, as expected for hinge-like or allosteric sites. A full implementation of this dynamic scoring is left for future work, but the definitions above are compatible with the present toolkit: any MD trajectory or ensemble can be processed by computing $\mathrm{Ep}_p(t_\ell)$ frame by frame and then applying user-chosen functionals of the discrete curve Γ_p.

Acknowledgements

This work was independently conceptualized and developed by the author (Haneul Park). The author utilized OpenAI ChatGPT models (GPT-4–class and successor models) to prototype analysis code, design statistical workflows (k-means clustering, distance-metric analysis, random-forest classification), polish LaTeX equations and notation, and refine the manuscript language. All AI-generated outputs were manually reviewed, validated, and adjusted by the author. All scientific reasoning, theoretical modeling, and experimental design originated from—and were confirmed by—the author.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[2] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Phil D. Adams, Caleb R. Glassman, Andy DeGiovanni, Jimin Pereira, Joao P. G. L. M. Rodrigues, Marc van Dijk, Andreas C. Ebrecht, Derren J. Opperman, Toni Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj Kumar Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[3] Predrag Radivojac, Wyatt T. Clark, Tal Ronnen Oron, and et al. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221–227, 2013.

[4] Vincent Le Guilloux, Peter Schmidtke, and Pierre Tuffery. fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, 10:168, 2009.

[5] John A. Capra, Roman A. Laskowski, Janet M. Thornton, Mona Singh, and Thomas A. Funkhouser. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3d structure. *PLoS Computational Biology*, 5(12):e1000585, 2009.

[6] Pablo Gainza, Freyr Sverrisson, Federico Monti, Emanuele Rodola, Davide Boscaini, Michael M. Bronstein, and Bruno E. Correia. Deciphering interaction fingerprints from protein surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

[7] Yu-Feng Lin, Chih-Wei Cheng, Chien-Sheng Shih, Jeng-Kuan Hwang, Chen-Shan Yu, and Chung-Hsien Lu. Mib: Metal ion-binding site prediction and docking server. *Journal of Chemical Information and Modeling*, 56(12):2287–2291, 2016.

[8] Liang Wei and Russ B. Altman. Recognizing protein binding sites using statistical descriptions of their 3d environments. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 497–508, 1999.

[9] Haim Ashkenazy, Elana Erez, Eric Martz, Tal Pupko, and Nir Ben-Tal. Consurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research*, 38(suppl_2):W529–W533, 2010.

[10] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Alexis Courbet, Robbert J. de Haas, Neville Bethel, Philip J. Y. Leung, Thomas F. Huddy, Samuel Pellock, Daniel Tischer, Franklin Chan, Brian Koepnick, Anh L. Nguyen, Aerin Kang, Bhanu Sankaran, Asim K. Bera, Neil P. King, and David Baker. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022.

[11] Weitao Yang and Willem J. Mortier. The use of global and local molecular parameters for the analysis of the gas-phase basicity of amines. *Journal of the American Chemical Society*, 108(19):5708–5711, 1986.

[12] Arieh Warshel, Pratul K. Sharma, Mitsunori Kato, Yi-Qin Xiang, Hua-Lin Liu, and Max M. Kamerlin. Electrostatic basis for enzyme catalysis. *Chemical Reviews*, 106(8):3210–3235, 2006.

[13] Wai Shing Tang, Gabriel Monteiro da Silva, Henry Kirveslahti, Erin Skeens, Bibo Feng, Timothy Sudijono, Kevin K. Yang, Sayan Mukherjee, Brenda Rubenstein, and Lorin Crawford. A topological data analytic approach for discovering biophysical signatures in protein dynamics. *PLoS Computational Biology*, 18(5):e1010045, 2022.

[14] Zixuan Cang and Guo-Wei Wei. Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology. *Bioinformatics*, 33:3549–3557, 2017.

[15] Berend Diepeveen, Shreyas Banerjee, Nikola B. Kovachki, Arnulf Jentzen, Luca Venturi, Kamyar Azizzadenesheli, Dusan Pavlovic, and Ozan Öktem. Riemannian geometry for efficient analysis of protein dynamics data. *Proceedings of the National Academy of Sciences of the United States of America*, 121(33):e2318951121, 2024.

[16] Huanwang Zheng, D. R. Cooper, P. J. Porebski, I. G. Shabalin, K. B. Handing, and W. Minor. Checkmymetal: a macromolecular metal-binding validation tool. *Acta Crystallographica Section D: Structural Biology*, 73(3):223–233, 2017.

[17] Paul D. Adams, Pavel V. Afonine, Gábor Bunkóczi, Vincent B. Chen, Ian W. Davis, Nathaniel Echols, Jeffrey J. Headd, Li-Wei Hung, Gary J. Kapral, Ralf W. Grosse-Kunstleve, Airlie J. McCoy, Nigel W. Moriarty, Robert Oeffner, Randy J. Read, Jane S. Richardson, David C. Richardson, Thomas C. Terwilliger, and Peter H. Zwart. Phenix: a comprehensive python-based system for macromolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography*, 66(2):213–221, 2010.

[18] Gerhard Langer, Stewart X. Cohen, Victor S. Lamzin, and Anastassis Perrakis. Automated macromolecular model building for x-ray crystallography using ARP/wARP version 7. *Nature Protocols*, 3(7):1171–1179, 2008.

[19] Pierre Hohenberg and Walter Kohn. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864–B871, 1964.

[20] Erich Runge and E. K. U. Gross. Density-functional theory for time-dependent systems. *Physical Review Letters*, 52(12):997–1000, 1984.

[21] Kelin Xia and Guo-Wei Wei. Persistent homology analysis of protein structure, flexibility, and folding. *Int. J. Numer. Meth. Biomed. Engng.*, 30(8):814–844, 2014.