WHERE DOES PROTEIN STRUCTURE EMERGE? LAYER-WISE NATIVE 3D ORGANIZATION ACROSS PROTEIN AND GENERAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein language models (PLMs) are trained primarily from amino-acid sequences, yet their representations support tasks closely tied to protein structure. We ask where, at what scale, and how consistently native three-dimensional organization becomes visible across network depth. We study dedicated PLMs spanning Transformer, convolutional, and state-space architectures, together with general-purpose language models and protein-oriented checkpoints. We analyze structure from two complementary perspectives: how the geometry of whole-protein representations changes across layers, and whether residues that are nearby in native 3D space remain nearby in representation space. We find strongly depth-dependent and model-specific organization. Whole-protein shape complexity often compresses across depth, but its trajectory varies across model families. In dedicated PLMs, Local Topological Fidelity (LTF) consistently peaks before the final layer, whereas general-purpose LLMs show more heterogeneous behavior. Secondary-structure conditioning further reveals that intermediate-scale neighborhoods ($k \approx 6$–9) centered on helical residues exhibit the strongest fidelity. Finally, LTF enables downstream-label-free, model-adaptive layer selection, with particularly strong utility on structure-sensitive evaluations. Across ten PLMs and four evaluations, it is the only downstream-label-free selector in the statistically top-ranked group alongside supervised layer-selection methods. The selected region remains stable across structural reference sets, supporting reuse without task-specific layer search.

1 INTRODUCTION

Protein function is inseparable from three-dimensional organization. Folding brings residues that may be distant along the sequence into physical proximity, forming catalytic sites, binding interfaces, and other functional motifs. Proteins with related structural organization can therefore retain similar functional properties even when their sequences differ substantially. Structure thus provides a natural bridge between amino-acid sequence and biological function.

PLMs are pretrained primarily from amino-acid sequences, without explicit protein structural information, yet their representations support tasks closely tied to protein structure, including contact and fold prediction Lin et al. (2023); Elnaggar et al. (2023); Ferruz et al. (2022); Ullanat et al. (2026); Bouras et al. (2026); Dill & MacCallum (2012); Schmirler et al. (2024); Rao et al. (2021). This suggests that sequence-only training can lead PLMs to encode aspects of native protein 3D organization in their latent representations. If such structural information is present, it should leave a measurable geometric footprint in representation space: residues that are neighbors in the folded protein may remain related in latent space, while whole-protein representations may exhibit systematic structural organization across layers. The resulting question is therefore not simply whether structural information is present, but *where, at what scale, and how consistently it becomes expressed across network depth*. A further practical question is whether layers in which native structural organization is more strongly expressed also provide better representations for downstream protein tasks. If so, structural fidelity could offer a way to choose representations without exhaustive task-specific layer search or downstream labels.

Existing PLM interpretability work has examined structural and biological information from several complementary perspectives, including attention and contact patterns, probing, coevolutionary

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

statistics, and sparse or concept-based features (Rao et al., 2020; Zhang et al., 2024; Vig et al., 2020; Simon & Zou, 2025; Gujral et al., 2025). Complementary representation-level studies have characterized the geometry of PLM embedding spaces through intrinsic dimensionality, neighborhood organization, embedding-space similarity, and direct comparison with native protein structure (Valeriani et al., 2022; 2023; Rissom et al., 2025; Beshkov & Malthe-Sørenssen, 2025; Vinod et al., 2026). Together, these studies establish that structural information is nontrivially reflected in PLM representations. What remains missing is a systematic account of how such organization varies across model families, scales, and training regimes, how recurrent geometric patterns relate to specific local structural elements, and when these patterns have practical downstream value.

To address these questions, we study layer-wise structural organization across dedicated PLMs spanning Transformer, convolutional, and state-space architectures, together with general-purpose language models processing protein sequences and protein-oriented checkpoints. Across this expanded model panel, we examine variation with model scale and training regime. We use two complementary geometric perspectives. Global Manifold Projection (GMP), summarized by intrinsic manifold dimensionality (IMD), characterizes how whole-protein representation-shape variability evolves across depth, whereas Local Topological Fidelity (LTF) measures how faithfully latent representations preserve native residue neighborhoods across multiple structural scales. We further condition LTF on residue-level secondary structure to resolve which local structural elements contribute to its depth- and scale-dependent patterns. We further test whether LTF can turn this structural analysis into a practical representation-selection signal: using an external structural reference set, it selects model-specific layers without target-task labels or exhaustive downstream layer search.

We find that native structural organization is strongly depth-dependent but does not follow a universal trajectory. Global geometric compression recurs across model families, yet its onset, persistence, and terminal behavior differ substantially with architecture and training regime. In dedicated PLMs, mean LTF consistently reaches its maximum before the final representation, although the preferred depth remains model-dependent; general-purpose LLMs instead exhibit more heterogeneous fidelity regimes. Secondary-structure conditioning further reveals that the recurring intermediate-neighborhood feature around $k \approx 6$–9 is strongly associated with helix-centered neighborhoods, while strand and coil exhibit different scale sensitivities. Finally, the same LTF signal supports downstream-label-free, model-adaptive layer selection for structure-sensitive PLM tasks. Across ten PLMs and four downstream evaluations, LTF is the only downstream-label-free selector in the statistically top-ranked group alongside supervised layer optimization. Fixed-depth comparisons show that its value lies in adapting to model-specific structural regions, while reference-set analyses show that these selections remain stable across datasets and can be estimated from a small reusable structural reference set.

2 Analyzing Structural Organization in Representation Space

A protein backbone can be represented as an ordered sequence of three-dimensional Cα points and, after interpolation, as a curve. At each model layer, the same residues become an ordered set of high-dimensional latent points that likewise defines a representation curve. This alignment lets us examine structure at two complementary levels: *Global Manifold Projection* (GMP) characterizes how whole-protein shape complexity changes across depth in the latent space, while *Local Topological Fidelity* (LTF) measures how faithfully native residue neighborhoods are preserved across structural scales. These global and local properties are related but not equivalent (Figure 1).

2.1 From Physical Coordinates to Residue Representations

For protein p with n_p analyzed residues, let $X_p = (x_{p,1}, \ldots, x_{p,n_p})^\top \in \mathbb{R}^{n_p \times 3}$ denote its native Cα coordinates and $H_p^{(\ell)} = (h_{p,1}^{(\ell)}, \ldots, h_{p,n_p}^{(\ell)})^\top \in \mathbb{R}^{n_p \times d_\ell}$ the corresponding layer-ℓ residue representations. The i-th row of both objects refers to the same residue, providing a physical anchor for asking how native structural organization in X_p is reflected in $H_p^{(\ell)}$ across depth.

We apply the same residue-aligned construction to dedicated PLMs, general-purpose LLMs processing protein sequences, and protein-oriented checkpoints. For general-purpose LLMs, amino acids are space-separated to facilitate residue-level extraction; prompt, boundary, and other non-residue tokens are excluded. InstructBioMol follows its native protein-sequence convention. Physical coordinates

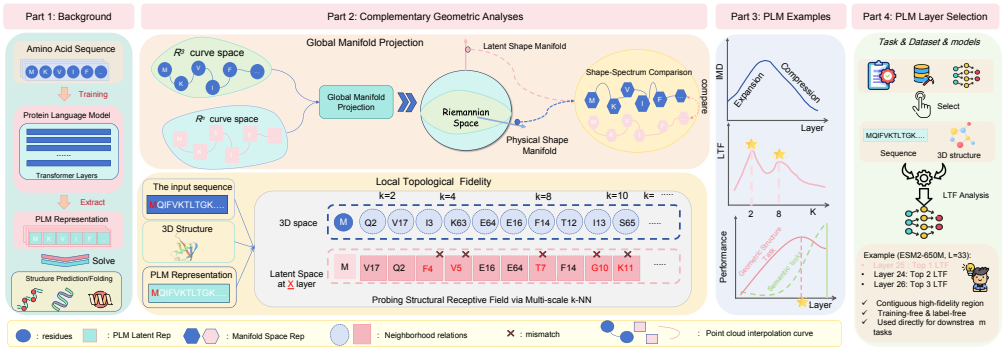

Figure 1: **Overview of the proposed analysis.** GMP tracks global whole-protein representation geometry, while LTF measures multi-scale preservation of native residue neighborhoods across layers. The resulting profiles characterize structural emergence and guide downstream-label-free PLM layer selection.

are used only as an external reference and are never provided to the sequence models. Implementation details are given in Appendices D.1 and C.

2.2 Global Shape Organization: GMP and IMD

Our first perspective asks how the *family of whole-protein representation shapes* changes across depth. Intuitively, if protein representation shapes vary along many independent geometric directions, their global organization is complex; if most variability can be described by only a few dominant modes, the shape family is comparatively compressed.

GMP operationalizes this view using Square-Root Velocity (SRV) shape analysis (Srivastava et al., 2010; Beshkov & Malthe-Sørenssen, 2025; Liu et al., 2010). Given an ordered protein representation, let $\mathbf{c} : [0, 1] \to \mathbb{R}^d$ denote the quadratic-spline interpolation of either the physical Cα coordinates ($d = 3$) or the layer-wise residue representations ($d = d_{\text{hidden}}$). We transform the resulting curve using its SRV representation,

$$\mathbf{q}(t) = \frac{\mathbf{c}'(t)}{\sqrt{\|\mathbf{c}'(t)\|_2}}, \tag{1}$$

which provides an elastic representation of curve shape. The complete pipeline—including interpolation, quotient shape-space operations, Karcher averaging, tangent-space linearization, and Tangent PCA—is detailed in Appendix B, with sampling-resolution details in Appendix B.6.

Within each curve family, we compute the Karcher mean and map the corresponding SRV representations to its tangent space through the Riemannian logarithm map. Let $\mathbf{v}_i^{(\ell)}$ denote the resulting tangent representation for protein i at layer ℓ. The tangent-space covariance is

$$C^{(\ell)} = \frac{1}{N-1} \sum_{i=1}^{N} \mathbf{v}_i^{(\ell)} \mathbf{v}_i^{(\ell)\top},$$

with eigenvalues $\{\lambda_k^{(\ell)}\}$. We summarize the concentration of this spectrum using the *Intrinsic Manifold Dimensionality* (IMD), defined as

$$\text{IMD}^{(\ell)} = \frac{\left(\sum_k \lambda_k^{(\ell)}\right)^2}{\sum_k \left(\lambda_k^{(\ell)}\right)^2}. \tag{2}$$

IMD measures the effective number of dominant geometric modes governing between-protein variation in whole-protein representation shapes. A high IMD indicates that many independent modes are needed to describe this variability, whereas a low IMD indicates concentration along fewer dominant modes. Accordingly, increasing IMD across layers reflects an expansion of global shape variability, while decreasing IMD indicates compression toward fewer recurring modes. GMP therefore provides a global, population-level view of representation geometry, distinct from whether individual residue neighborhoods preserve their native physical organization.

The same GMP pipeline applied to the physical Cα structures yields a reference value of $\mathrm{IMD}_{\mathrm{phys}} \approx 3.88$. This value characterizes the effective dimensionality of variation across protein shapes rather than the three Cartesian coordinates of an individual structure. Importantly, physical and latent curves are analyzed within their respective shape spaces; GMP does not assume a pointwise isometric embedding between $\mathbb{R}^3$ and the high-dimensional latent space. The dimensionless IMD statistic instead provides a commensurable summary of their global shape complexity. Consequently, proximity to $\mathrm{IMD}_{\mathrm{phys}}$ is descriptive of global geometric organization and does not, by itself, establish preservation of native residue neighborhoods.

2.3 Local Topological Fidelity (LTF)

Global shape complexity does not reveal whether a representation preserves the native local organization of an individual protein. LTF therefore asks a more direct question: *for each residue, do the residues that are nearby in native 3D space also remain nearby in the latent representation?*

For a protein with residues $\{r_1, \ldots, r_n\}$, let $A^{(k)}_{\mathrm{phys}} \in \{0,1\}^{n\times n}$ denote the directed binary k-nearest-neighbor adjacency matrix constructed from Euclidean distances between Cα coordinates, and let $A^{(k)}_{\mathrm{latent},\ell}$ denote the corresponding graph constructed from layer-ℓ latent representations using cosine distance. The two graphs share the same residue identities and exclude self-edges, allowing their neighborhood structure to be compared directly. We evaluate $k \in \{1, \ldots, 19\}$, where k denotes the number of neighbors rather than a physical-distance or sequence-separation threshold. Replacing cosine distance with Euclidean distance in latent space yields identical or adjacent selected layers across the evaluated PLMs (Appendix F.8).

To provide a chance-level reference, we additionally construct $A^{(k)}_{\mathrm{null}}$ from an isotropic Gaussian embedding of matching dimensionality. Sensitivity to the choice of null model is characterized in Appendix F.7. The score at layer ℓ and neighborhood scale k is

$$S^{(\ell)}_{\mathrm{LTF}}(k) = \frac{\mathbb{E}_i\left[\left\|A^{(k)}_{\mathrm{phys},i} - A^{(k)}_{\mathrm{null},i}\right\|_F\right]}{\mathbb{E}_i\left[\left\|A^{(k)}_{\mathrm{phys},i} - A^{(k)}_{\mathrm{latent},i,\ell}\right\|_F\right]}, \tag{3}$$

where the expectation is taken over proteins i in the structural reference set. Higher S_{LTF} indicates that the latent representation preserves native physical neighborhoods more faithfully relative to the null reference.

Varying k exposes the scale dependence of this correspondence. Small neighborhoods emphasize local backbone and covalent organization, whereas larger neighborhoods increasingly capture secondary- and tertiary-scale spatial relationships. We refer to the resulting variation of LTF over k as the model's multi-scale structural profile. The exact LTF computation protocol is provided in Appendix D.2.

Secondary-structure-conditioned LTF. To resolve which local structural elements contribute to the LTF signal, we assign each aligned residue to the standard three-state secondary structure (SS3) grouping using DSSP: helix ($H/G/I$), strand (E/B), or coil ($T/S/C/-$). Each complete protein sequence is processed unchanged, and the physical, latent, and null k-NN graphs are constructed over all aligned residues exactly as in the standard LTF pipeline.

Conditioning is applied only during aggregation. After computing residue-level LTF, we average scores over anchor residues belonging to each SS3 category, while candidate neighbors remain drawn from the entire protein. Thus, Helix-LTF, Strand-LTF, and Coil-LTF measure the fidelity of neighborhoods *centered on* different secondary-structure elements, rather than fidelity within isolated secondary-structure fragments. A formal residue-level definition is provided in Appendix D.3

GMP and LTF therefore characterize two distinct levels of structural organization. GMP describes how whole-protein shape variability evolves across the representation hierarchy, whereas LTF measures how faithfully native within-protein neighborhood relationships are preserved across depth and structural scale. With these quantities defined, we now turn to the central empirical question of the paper: *where does native 3D organization emerge across model depth, and how consistently does it appear across model families and training regimes?*

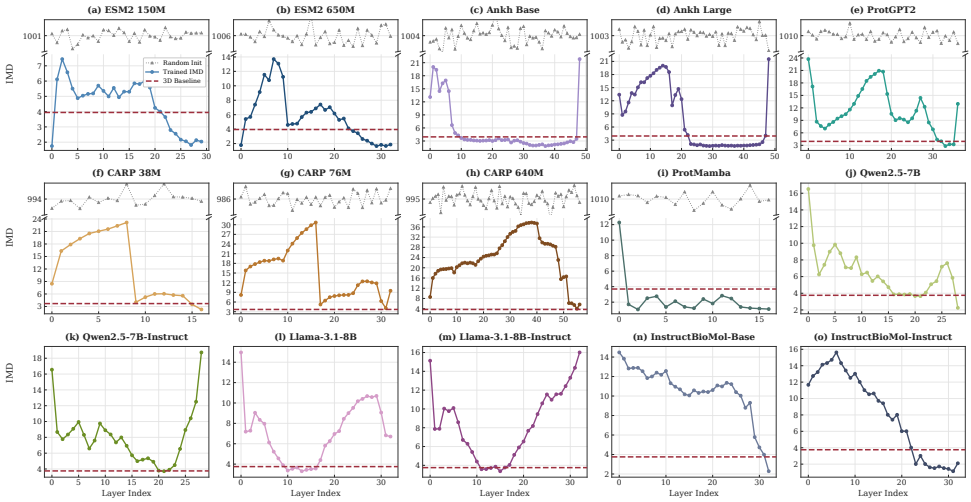

Figure 2: Layer-wise Intrinsic Manifold Dimensionality (IMD) is shown for all 15 analyzed models.

3 Where Does Native 3D Organization Emerge?

With GMP and LTF defined, we now ask how native structural organization varies across model depth, architectures, and training regimes. We analyze protein domains from SCOPe (Chandonia et al., 2022) across dedicated PLMs—ESM2 Lin et al. (2023), Ankh Elnaggar et al. (2023), CARP Yang et al. (2024), ProtGPT2 Ferruz et al. (2022), and ProtMamba Sgarbossa et al. (2025)—together with Base/Instruct Qwen2.5-7B and Llama-3.1-8B, and protein-oriented InstructBioMol Zhuang et al. (2025) checkpoints across training stages. Exact checkpoint coverage for GMP and LTF is shown in the corresponding figures. We examine which structural patterns recur across this panel and where they differ, focusing on global shape complexity, local fidelity across depth and scale, training-regime variation, and secondary-structure-conditioned organization.

3.1 Global Geometric Compression Recurs, but Its Trajectory Is Model-Dependent

Previous analyses of protein and general language models have reported an expansion–compression pattern in representation dimensionality across depth (Beshkov & Malthe-Sørenssen, 2025; Valeriani et al., 2023). Our broader panel recovers this behavior in several families but also reveals an important qualification: *geometric compression is recurrent, whereas a single expansion–compression trajectory is not*. As shown in Figure 2, substantial IMD reductions occur across the trained models, but their onset, persistence, and terminal behavior differ markedly.

The morphology of this reorganization is strongly model-dependent. ESM2 most clearly follows early expansion followed by later compression, whereas CARP exhibits a more extended high-dimensional regime and a sharper, step-like collapse. Other families depart further from this two-stage pattern: Ankh enters a prolonged low-dimensional regime before terminal re-expansion, ProtGPT2 undergoes repeated contraction and re-expansion, and ProtMamba compresses early and remains low-dimensional through most of its depth. Randomly initialized controls remain at substantially higher IMD and do not reproduce these low-dimensional regimes, indicating that the observed trajectories arise through learned representations rather than initialization alone.

General-purpose LLMs provide an additional test of which patterns are protein-specific. Qwen2.5 and Llama-3.1 both exhibit intermediate compression followed by later re-expansion, consistent with non-monotonic dimensional dynamics reported in general Transformers (Valeriani et al., 2023), whereas protein-oriented InstructBioMol shows more sustained late compression with checkpoint-dependent variation. The presence of pronounced compression in general-purpose LLMs is consequential: global compression alone cannot be interpreted as native protein-structure alignment. The dimensionless IMD statistic therefore provides a global view of how whole-protein shape complexity reorganizes across latent layers. To complement this population-level perspective, we next turn to LTF, which asks locally whether native residue neighborhoods are preserved in the representation space.

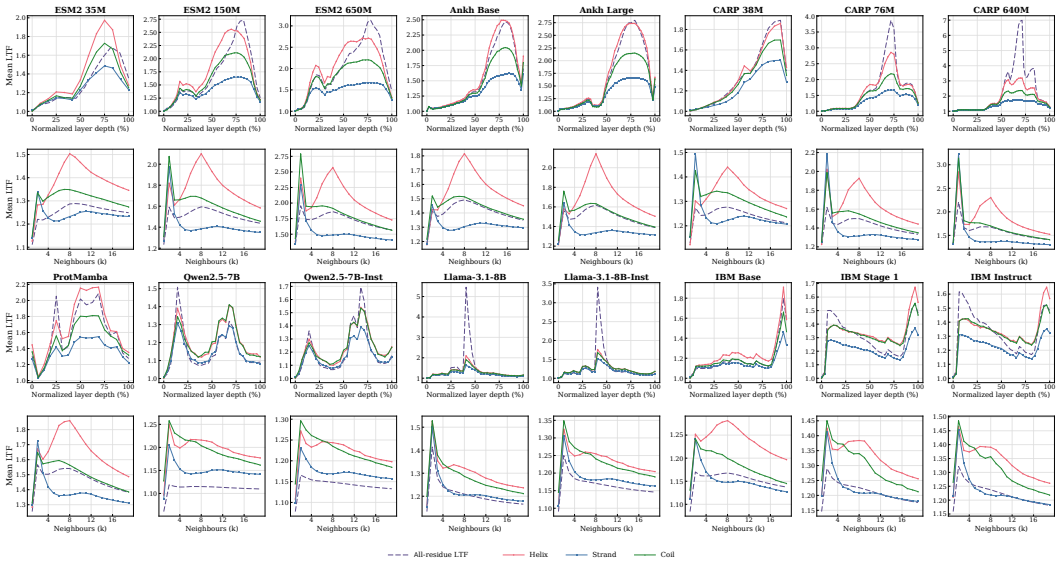

Figure 3: Complete LTF profiles across depth, neighborhood scale, and secondary-structure categories.

3.2 Native-Neighborhood Fidelity Forms Distinct Depth and Scale Regimes

LTF complements the global IMD analysis along two axes: network depth asks *where* native-neighborhood fidelity is strongest, while neighborhood size k asks *at what structural scale* it is expressed.

Across every dedicated PLM evaluated, the strongest all-residue LTF appears before the final layer and exceeds final-layer fidelity. These maxima generally lie in a mid-to-late region, but their morphology is model-dependent: ESM2 and Ankh form broad high-fidelity regions, CARP exhibits sharper peaks—with CARP-38M much closer to the output—and ProtMamba contains multiple elevated regions. Model scale mainly changes the strength and shape of these regimes rather than imposing a consistent depth shift; larger ESM2 and CARP variants generally show stronger fidelity, but the response remains family-specific. The recurring pattern is therefore a *pre-final preference*, not a universal optimal depth.

The scale profiles reveal a second recurring organization. Many dedicated PLMs show both a very-local regime at small $k \approx 2$ and a broader intermediate-scale elevation around $k \approx 6$–9. This multi-scale structure recurs across Transformer, convolutional, and state-space PLMs, although the relative prominence of the two regimes varies by model. Thus, the layer at which native structure is most strongly expressed and the structural scale at which it is preserved constitute related but distinct properties of the representation.

General-purpose LLMs differ along both axes. Their all-residue LTF remains above the null reference, indicating measurable native-neighborhood alignment, but their depth profiles are substantially more heterogeneous: Qwen2.5 shows separated early and later regimes, whereas Llama-3.1 concentrates its strongest alignment in a narrow intermediate region. Their scale profiles are also more dominated by the smallest neighborhoods and show weaker intermediate-scale structure than most dedicated PLMs. Protein-oriented InstructBioMol occupies an intermediate regime, combining checkpoint-dependent depth profiles with more visible meso-scale organization. We next examine how these regimes shift across training stages and which local structural elements underlie the intermediate-scale feature.

3.3 Structural Fidelity Shifts Across Training Regimes

Across related checkpoints, training-regime differences reorganize native-neighborhood fidelity in a model-dependent rather than uniform manner. Qwen2.5 shifts from a stronger early regime in Base toward a more prominent later peak after instruction tuning, whereas Llama-3.1 largely preserves the location of its sharp intermediate peak while reducing its magnitude. InstructBioMol shows a third pattern: Stage 1 and Instruct strengthen early fidelity relative to Base while retaining a late

resurgence. Thus, post-training is associated primarily with a redistribution of *where* structural fidelity is expressed across depth, rather than a consistent increase in fidelity.

The global IMD profiles do not necessarily follow these local changes. Qwen2.5 and Llama-3.1 both show stronger terminal IMD re-expansion in their instruction-tuned checkpoints, yet their LTF profiles change in different ways. Global representation-shape complexity and local native-neighborhood fidelity therefore respond differently across training regimes, reinforcing that they capture distinct aspects of representation organization. We next examine which local structural elements underlie the scale-dependent LTF patterns.

3.4 Secondary-Structure Conditioning Reveals Scale- and Depth-Specific Organization

The recurring intermediate-scale LTF feature raises a natural question: which local structural elements underlie this pattern, and how do their fidelity profiles evolve across depth? We address this using the SS3-conditioned LTF defined in Section 2.3.

The clearest distinction appears across neighborhood scale. Across the dedicated PLMs, neighborhoods centered on helical residues exhibit a pronounced LTF elevation around $k \approx 6$–9, consistently across ESM2, Ankh, CARP, and ProtMamba. Strand-centered fidelity is more concentrated at the smallest neighborhoods, whereas coil shows a broader and more moderate intermediate-scale response. This residue-level analysis provides a more direct interpretation of the previously reported meso-scale feature near $k \approx 8$ (Beshkov & Malthe-Sørenssen, 2025): it is strongly associated with neighborhoods centered on helical residues, rather than being an undifferentiated property of all residue types. Here, k denotes neighborhood size rather than helix length, and candidate neighbors are not restricted to the anchor's secondary- structure category.

The depth profiles provide a complementary view. In most dedicated PLMs, helix-, strand-, and coil-centered fidelity rise within the same broad mid-to-late depth region as the all-residue LTF, indicating that the pre-final regime is shared across multiple secondary-structure contexts rather than isolated to a single residue class. Helix-centered fidelity is often the most pronounced within this region, consistent with its distinctive meso-scale signature. The correspondence is not exact, however: several models exhibit sharper or slightly shifted all-residue peaks, while the narrow intermediate peaks of general-purpose LLMs are more prominent in the all-residue profile than in any single SS3-conditioned curve. Secondary-structure conditioning therefore identifies *where part of the local structural signal is organized and at what scale*, but does not provide an additive decomposition of all-residue LTF. General-purpose LLMs further emphasize very-local neighborhoods, whereas protein-oriented InstructBioMol retains more visible helix-associated intermediate-scale organization. Overall, secondary-structure effects are expressed more consistently through the *strength and scale* of native-neighborhood preservation than through a universal ordering of emergence depth.

4 When Is Structural Fidelity Useful?

Section 3 shows that native-neighborhood fidelity is strongly depth-dependent in dedicated PLMs, with mean LTF consistently peaking before the final representation but at model-dependent depths. We now ask whether this descriptive signal has a practical consequence: *can structural fidelity identify useful representation layers without downstream labels?* To test this, we select the layer with the highest LTF averaged across neighborhood scales and evaluate the resulting representations on structure-sensitive downstream tasks. We restrict this analysis to dedicated PLMs, separating the layer-selection application from the descriptive general-LLM results above.

4.1 From LTF to Downstream-Label-Free PLM Layer Selection

We summarize each layer by its mean fidelity across the neighborhood scales analyzed in Section 2.3:

$$\mathcal{S}(\ell) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} S^{(\ell)}_{\mathrm{LTF}}(k), \qquad \ell^*_{\mathrm{LTF}} = \operatorname*{arg\,max}_{\ell \in \{1,\ldots,L\}} \mathcal{S}(\ell), \tag{4}$$

where $\mathcal{K} = \{1, \ldots, 19\}$. Thus, the selected layer is the one that most consistently preserves native physical neighborhoods across structural scales. This is the same mean-LTF quantity used in the layer-wise analyses of Section 3.

We compute the selector on protein domains from SCOPe and evaluate the resulting frozen representations on CATH contact/distance prediction, SCOPe fold classification, and ECPred enzyme classification. Selection requires known structures but no downstream labels, gradients, or parameter updates; downstream probes are trained only after the layer has been selected. Detailed dataset, probe, and training specifications are provided in Appendices C, D.4, and D.5.

4.2 LTF-Selected Layers Improve Structure-Sensitive Representation Utility

Across the dedicated PLM panel, layers selected by mean LTF provide consistently stronger representations for structure-sensitive prediction. For contact/distance prediction, the LTF-selected layer achieves lower MAE than the final representation across all ten PLMs, reducing the mean from 9.10 to 8.52 Å. Table 1 illustrates this behavior on ESM2-650M: LTF selects layer 25 and reduces contact MAE from 7.97 to 7.29 Å, while improving fold accuracy from 0.72 to 0.76 and fine-grained EC-L4 accuracy from 0.47 to 0.49. In contrast, coarse EC-L1 slightly favors the final layer (0.77 versus 0.76), already indicating that the value of structural fidelity depends on the downstream task.

Table 1: Representative comparison on ESM2-650M (L=33), anchoring the key comparisons in the text. Contact prediction on CATH 4.2; fold classification on SCOPe; EC classification on ECPred. Bold: best unsupervised per metric. ‡ Unsupervised, structure-aware. † Supervised.

		Contact Prediction			Fold	EC Classification	
Method	ℓ	P@L	P@$L/2$	MAE (Å)	Acc.	L1 Acc.	L4 Acc.
Random (Uniform)	–	.40±.08	.43±.03	9.06±.5	.61±.04	.66±.02	.38±.03
Random (Top-50%)	–	.47±.07	.56±.09	8.40±.6	.69±.03	.70±.02	.42±.03
PRC-Selected‡	6	.38	.42	9.50	.57	.62	.34
Final Layer	33	.50	.66	7.97	.72	**.77**	.47
Fixed Depth (80%)	26	.50	.67	7.45	.74	.72	.46
Mean Pool	–	.42	.53	8.80	.70	.72	.41
IMD-Selected‡	23	.52	.58	8.75	.73	.73	.43
LTF-Selected‡	**25**	.51	**.68**	**7.29**	**.76**	.76	**.49**
Learned Wt. Sum†	–	.55	.68	7.40	.78	.77	.49
Grid Search Oracle†	–	.58	.69	7.29	.79	.77	.49

The result extends beyond a single model or task. Table 2 summarizes the aggregate comparison across the pool of ten selectors, ten PLMs, and four downstream evaluations. LTF attains a mean rank of 2.73, compared with 5.72 for Fixed 80%, 5.83 for IMD selection, and 6.73 for the final layer. A Friedman test rejects equal rankings ($p < 0.001$), and LTF is the only downstream-label-free method in the top Nemenyi group alongside the supervised learned-weighted-sum and grid-search oracle. Removing the in-domain fold task leaves the ranking highly stable: LTF changes from 2.73 to 2.69 and remains the only downstream-label-free method in the top group. Thus, the layer-wise native-neighborhood signal identified in Section 3 is associated with downstream utility across model families rather than only on the representative ESM2 setting.

Table 2: Aggregated mean rank (↓) across 10 models × 4 tasks ($N = 40$ evaluations), 10 methods. Friedman test: $p < 0.001$ (df $= 9$, $N = 40$). CD $= 2.14$ (Nemenyi, $\alpha = 0.05$). Full results in App. E.3, Table 12.

Method	**Mean Rank**	**Group**
Grid Search Oracle†	1.23	A
Learned Weighted Sum†	2.12	A
LTF-Selected	**2.73**	**A**
Fixed Depth (80%)	5.72	B
IMD-Selected	5.83	B
Mean Pool	6.23	BC
Final Layer	6.73	BC
PRC-Selected	7.36	BCD
Random (Top-50%)	8.01	CD
Random (Uniform)	9.04	D

Importantly, not every structure-aware geometric criterion is equally useful for selection. IMD-based selection and global pairwise distance correlation perform substantially below LTF in the aggregate comparison (Table 2). This distinction mirrors the analysis in Section 2: global shape complexity characterizes how representation geometry evolves, whereas LTF directly measures whether native local neighborhoods are preserved. The latter therefore provides the more operational signal for selecting representations when downstream prediction depends on protein structural organization.

4.3 Model-Adaptive Selection Beyond Fixed-Depth Heuristics

A natural alternative to LTF is to use the same proportional depth for every model. We therefore sweep fixed layers from 60% to 90% of network depth (Table 3). The strongest fixed rule by mean contact MAE is 75% depth (8.61 Å), close to the 8.52 Å obtained by LTF.

The practical distinction is clearer in model adaptivity and reliability: the lowest observed worst-case regret among the tested fixed rules is 0.54 Å, compared with 0.20 Å for LTF. The reason is that the preferred depth varies across models. For ESM2-650M, the 75% and 80% rules bracket the LTF-selected region, whereas 70% is the strongest tested fixed choice for CARP-640M but performs substantially worse on ESM2-650M. CARP-38M favors an even later representation beyond the

tested 90% depth. Thus, the practical value of LTF is not a large average advantage over every tuned fixed-depth heuristic, but model-adaptive selection with lower observed worst-case failure without choosing a depth fraction from downstream labels. Sweep summaries and representative model-wise results are reported in Appendix F.1.

4.4 Reference-Set Transferability and Robustness

A practical layer selector should not require a new structural reference set for every downstream application. We therefore recompute LTF using structural subsets from CATH, ECPred, and ProteinGym and compare the resulting profiles with the default SCOPe reference (Table 4). On ESM2-650M, the selected layer shifts by at most one position from the SCOPe-selected layer 25, while the full layer-wise profiles remain correlated ($\rho = 0.71$–0.76) and four of the top five layers are retained. Downstream performance changes by less than 0.5%. LTF therefore identifies a stable model-specific structural region rather than a reference-set-specific optimum, supporting reuse of a generic structural reference set across downstream applications. This practical requirement is further reduced by reference-size robustness: layer selection is already near-converged with approximately 100 diverse proteins. Complementary controls show that the selected region is also stable to sequence-local filtering, null-model choice, and latent distance metric. Native-reference interventions further yield a substantially larger native-versus-rewired LTF difference at the LTF-selected layer than at the final layer (36.5% versus 11.5%), while representation perturbations show the same concentration (GSI 1.86 versus 1.18). Together, these results indicate that LTF can be estimated from a relatively small, reusable structural reference set while remaining robust to major analysis choices. Full robustness analyses are provided in Appendix F.

Table 3: Fixed-depth sweep. Contact MAE is averaged over ten PLMs. Worst regret is the largest model-wise MAE gap to the supervised oracle. † Supervised reference.

Selector	Mean MAE (Å) ↓	Worst Regret (Å) ↓
Fixed 60%	9.02	1.34
Fixed 65%	8.92	1.39
Fixed 70%	8.78	1.49
Fixed 75%	8.61	0.59
Fixed 80%	8.70	0.54
Fixed 85%	8.92	0.77
Fixed 90%	9.06	1.00
LTF-Selected	**8.52**	**0.20**
Grid Search Oracle†	8.33	0.00

5 Related Work

Our work connects three lines of research. PLM interpretability studies biological information encoded in hidden states through attention, probing, and sparse or concept-based analyses (Vig et al., 2020; Rao et al., 2020; Simon & Zou, 2025; Gujral et al., 2025). Geometric and topological work characterizes intrinsic dimensionality and representation shape (Ansuini et al., 2019; Valeriani et al., 2023; Srivastava et al., 2010); recent PLM studies further connect such geometry to protein structure (Beshkov & Malthe-Sørenssen, 2025), but systematic global–local comparisons across model families remain limited. Layer-wise probing motivates representation selection but is typically supervised and task-specific (Belinkov, 2021; Jawahar et al., 2019; Skean et al., 2025; Pinto, 2025). We instead combine global whole-protein shape analysis with local native-neighborhood fidelity, using physical structures as an external anchor, and show that the latter supports downstream-label-free layer selection. Full discussion appears in Appendix A.

Table 4: Reference-set transferability on ESM2-650M (L=33). LTF is recomputed independently on each structural reference set. Profile ρ denotes Spearman correlation with the SCOPe-derived layer-wise LTF profile; Top-5 is the overlap of the five highest-ranked layers.

Reference Set	ℓ^*	Profile ρ	Top-5	ΔPerf
SCOPe (default)	25	–	5/5	–
CATH 4.2	24	.76	4/5	$< 0.5\%$
ECPred subset	26	.72	4/5	$< 0.3\%$
ProteinGym subset	25	.71	4/5	$< 0.2\%$

6 Conclusion

We investigated where and at what scale native 3D organization emerges across protein and general language models. Global and local analyses reveal depth-dependent, model-specific organization: geometric compression recurs without a universal trajectory, while dedicated PLMs consistently exhibit strongest native-neighborhood fidelity before the final layer. Secondary-structure conditioning further resolves the structural scales underlying this organization, and LTF turns the local signal into downstream-label-free, model-adaptive layer selection for structure-sensitive tasks. Broader implications, limitations, and future directions are discussed in Appendix H.

AI USE STATEMENT

Generative AI tools were used only for grammar checking, language polishing, and limited assistance with manuscript organization. They were not used to formulate the scientific hypotheses, define the methods, determine the experimental design, or interpret the reported results. All AI-assisted text was reviewed and revised by the authors, who take full responsibility for the scientific content of this work.

ETHICS STATEMENT

This work uses publicly available protein sequence, structure, and functional datasets and involves no human subjects, personal data, or personally identifiable information. All experiments are computational and focus on analyzing representations of pretrained protein and general-purpose language models. We perform no protein synthesis, experimental validation, or biological deployment, and the reported computational analyses should not be interpreted as experimentally validated biological conclusions.

REPRODUCIBILITY STATEMENT

The main methodology and experimental design are described in Sections 2–4. Detailed GMP and IMD procedures are provided in Appendix B; datasets and model coverage in Appendix C; representation extraction, LTF computation, probing protocols, hyperparameters, and statistical procedures in Appendix D; and complete downstream, robustness, and exploratory analyses in Appendices E–G. Code, configurations, and supporting experimental artifacts will be released upon acceptance to facilitate reproduction of the reported results.

REFERENCES

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48:207–219, 2021. doi: 10.1162/coli_a_00422.

Kosio Beshkov and Anders Malthe-Sørenssen. Towards understanding the shape of representations in protein language models. *arXiv preprint arXiv:2509.24895*, 2025.

Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Simsekli. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in neural information processing systems*, 34:6776–6789, 2021.

George Bouras, Susanna R Grigson, Milot Mirdita, Michael Heinzinger, Bhavya Papudeshi, Vijini Mallawaarachchi, Renee Green, Rachel Seongeun Kim, Victor Mihalia, Alkis James Psaltis, et al. Protein structure-informed bacteriophage genome annotation with phold. *Nucleic Acids Research*, 54(1):gkaf1448, 2026.

John-Marc Chandonia, Lindsey Guan, Shiangyi Lin, Changhua Yu, Naomi K Fox, and Steven E Brenner. Scope: improvements to the structural classification of proteins–extended database to facilitate variant interpretation and machine learning. *Nucleic acids research*, 50(D1):D553–D559, 2022.

Ken A Dill and Justin L MacCallum. The protein-folding problem, 50 years on. *science*, 338(6110): 1042–1046, 2012.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized protein language model unlocks general-purpose modelling. *arXiv preprint arXiv:2301.06568*, 2023.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Onkar Gujral, Mihir Bafna, Eric Alm, and Bonnie Berger. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122(34):e2506316122, 2025.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, 20(1):723, 2019.

Magnus Haraldson Høie, Erik Nicolas Kiehl, Bent Petersen, Morten Nielsen, Ole Winther, Henrik Nielsen, Jeppe Hallgren, and Paolo Marcatili. Netsurfp-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic acids research*, 50(W1): W510–W515, 2022.

Chao Hou, Di Liu, and Yufeng Shen. Unsupervised discovery of functional sequence patterns from protein language model with motifae. *Nature Communications*, 2026.

Andrea Hunklinger and Noelia Ferruz. Toward the explainability of protein language models. *arXiv preprint arXiv:2506.19532*, 2025.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 3651–3657, 2019.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Wei Liu, Anuj Srivastava, and Jinfeng Zhang. Protein structure alignment using elastic shape analysis. In *Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology*, pp. 62–70, 2010.

Gregory Naitzat, A. Zhitnikov, and Lek-Heng Lim. Topology of deep neural networks. *ArXiv*, abs/2004.06093, 2020.

Luis Pinto. Superior molecular representations from intermediate encoder layers. *arXiv preprint arXiv:2506.06443*, 2025.

Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *Biorxiv*, pp. 2020–12, 2020.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International conference on machine learning*, pp. 8844–8856. PMLR, 2021.

Pia Francesca Rissom, Paulo Yanez Sarmiento, Jordan Safer, Connor W. Coley, Bernhard Y. Renard, H. Heyne, and Sumaiya Iqbal. Decoding protein language models: insights from embedding space analysis. *bioRxiv*, 2025. doi: 10.1101/2024.06.21.600139.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the national academy of sciences*, 118(15):e2016239118, 2021.

Robert Schmirler, M. Heinzinger, and B. Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15, 2024. doi: 10.1038/s41467-024-51844-2.

Siddharth Setlur, Djordje Mihajlovic, and Darrick Lee. Interpreting latent protein language model features with geometric annotations. *arXiv preprint arXiv:2608.26419*, 2026.

Damiano Sgarbossa, Cyril Malbranke, and Anne-Florence Bitbol. Protmamba: a homology-aware but alignment-free protein state space model. *Bioinformatics*, 41(6), 2025.

Aleena Siji, Amir Mohammad Karimi Mamaghan, Ferdinand Kapl, Tobias Höppe, Emmanouil Angelis, Andrea Dittadi, Maurice Brenner, Michael Heinzinger, Karl Henrik Johansson, Kaitlin Maile, Johannes von Oswald, and Stefan Bauer. From words to amino acids: Does the curse of depth persist?, 2026. URL https://arxiv.org/abs/2602.21750.

Elana Simon and James Zou. Interplm: discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, pp. 1–11, 2025.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*, 2025.

Anuj Srivastava, Eric Klassen, Shantanu H Joshi, and Ian H Jermyn. Shape analysis of elastic curves in euclidean spaces. *IEEE transactions on pattern analysis and machine intelligence*, 33(7): 1415–1428, 2010.

Jacob L Steenwyk. A geometric atlas of how esm3 organizes modalities across depth. *bioRxiv*, pp. 2026–07, 2026.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *BioRxiv*, pp. 2023–10, 2023.

Varun Ullanat, Bowen Jing, Samuel Sledzieski, and Bonnie Berger. Learning the language of protein-protein interactions. *Nature Communications*, 2026.

L. Valeriani, F. Cuturello, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of protein language models. *bioRxiv*, 2022. doi: 10.1101/2022.10.24.513504.

Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36:51234–51252, 2023.

Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222*, 2020.

Ria Vinod, Ava P Amini, Lorin Crawford, and Kevin K Yang. Trainable subnetworks reveal insights into structure knowledge organization in protein language models. *PLOS Computational Biology*, 22(2):e1013925, 2026.

Shawn T Whitfield, Tom Marty, Robert M Vernon, Christopher James Langmead, Dhanya Sridhar, and Quentin Fournier. High-resolution dissection of concept acquisition in different families of protein language models. *bioRxiv*, pp. 2026–07, 2026.

Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in neural information processing systems*, 34:4738–4750, 2021.

Kevin K Yang, Nicolo Fusi, and Alex X Lu. Convolutions are competitive with transformers for protein sequence pretraining. *Cell systems*, 15(3):286–294, 2024.

Zhidian Zhang, Hannah K Wayment-Steele, Garyk Brixi, Haobo Wang, Dorothee Kern, and Sergey Ovchinnikov. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences*, 121(45):e2406285121, 2024.

Xiang Zhuang, Keyan Ding, Tianwen Lyu, Yinuo Jiang, Xiaotong Li, Zhuoyi Xiang, Zeyuan Wang, Ming Qin, Kehua Feng, Jike Wang, Qiang Zhang, and Huajun Chen. Advancing biomolecular understanding and design following human instructions. *Nature Machine Intelligence*, 7(7): 1154–1167, 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01064-0. URL http://dx.doi.org/10.1038/s42256-025-01064-0.

APPENDIX: CONTENTS

A Extended Related Work 16
A.1 PLM Interpretability: From Attention to Sparse Features 16
A.2 Layer-Wise Organization of Biological Knowledge 16
A.3 Representation Geometry, Topology, and Native Structure 16
A.4 Layer Selection and Depth-Dependent Representation Utility 17
A.5 Positioning of the Present Work 17
B GMP and IMD Technical Details 18
B.1 Discrete-to-Continuous Interpolation 18
B.2 Square-Root Velocity Representation 18
B.3 Shape-Space Alignment 18
B.4 Karcher Mean and Tangent-Space Analysis 18
B.5 Intrinsic Manifold Dimensionality 18
B.6 Sampling Resolution 19
C Dataset and Model Details 20
C.1 SCOPe Structural Reference Set 20
C.2 CATH 4.2 Contact and Distance Prediction 20
C.3 ECPred Enzyme Classification 20
C.4 ProteinGym Exploratory Mutation Evaluation 20
C.5 DeepLoc Applicability Analysis 21
C.6 Models and Analysis Coverage 21
D Experimental Protocols 22
D.1 Representation Extraction 22
D.2 LTF Computation and Mean-over-Scale Selection 22
D.3 Secondary-Structure-Conditioned LTF 22
D.4 Probing Heads and Task Metrics 23
D.5 Training Hyperparameters 23
D.6 Definitions of the Ten Comparison Methods 23
D.7 Statistical Testing Protocols 24
D.8 Computational Budget 24
E Downstream Results and Statistical Analyses 25
E.1 Contact Prediction across Ten PLMs 25
E.2 EC Accuracy and Macro-F1 25
E.3 Aggregate Statistical Comparison 25
E.4 Aggregate Excluding In-Domain Fold Classification 26
F Layer-Selection Robustness and Diagnostic Controls 27
F.1 Fixed-Depth Sweep 27
F.2 Simple Neighborhood Selectors and LTF Components 27
F.3 Selected Structural Region 29
F.4 Reference-Set Robustness 29
F.5 Sequence-Locality Controls 30
F.6 Native-Structure Specificity 31
F.7 Null-Model Sensitivity 31
F.8 Latent Distance Metric 31
F.9 Pairwise-Distance Diagnostic 32
F.10 Probe Hyperparameters across Architectures 34
F.11 Representation Perturbation Analysis 34
F.12 Task Dependence: DeepLoc 36
G Mutation Effect Prediction: Exploratory Analysis 37
G.1 Zero-Shot Scoring Protocol 37
G.2 Assay Subset 37
G.3 Five-Model Results 37
G.4 Per-Assay Results 38

H Limitations, Future Directions, and Broader Impact 39
H.1 Practical Implications 39
H.2 Limitations 39
H.3 Future Directions 39
H.4 Broader Impact 40

A Extended Related Work

Our work lies at the intersection of protein-language-model interpretability, representation geometry, layer-wise analysis, and representation selection. These directions address related but distinct questions: what biological information is encoded, where that information appears across depth, how the representation space itself is organized, and whether such organization can guide downstream use. We review these connections separately before positioning the present study.

A.1 PLM Interpretability: From Attention to Sparse Features

A broad literature has established that protein language models encode biologically meaningful information despite being trained primarily from sequence. Early interpretability studies examined attention patterns and showed that they contain signals related to residue contacts, binding sites, and evolutionary couplings (Vig et al., 2020; Rao et al., 2020; Zhang et al., 2024). Probe-based analyses similarly demonstrate that frozen PLM representations support prediction of secondary structure, solvent accessibility, remote homology, and other biological properties (Rives et al., 2021; Elnaggar et al., 2021; Heinzinger et al., 2019; Høie et al., 2022).

More recent work moves from decoding predefined labels toward identifying interpretable internal features. Sparse-autoencoder approaches such as InterPLM and related studies recover latent features associated with structural motifs, functional regions, and biological annotations (Simon & Zou, 2025; Gujral et al., 2025). MotifAE further introduces local smoothness into sparse feature discovery and links latent features to functional motifs, structural domains, and folding stability (Hou et al., 2026). Geometric annotation of sparse PLM features provides a complementary residue-level view by associating SAE features with descriptors of the Cα backbone (Setlur et al., 2026). Trainable-subnetwork analyses instead intervene on subsets of model parameters and show that structural categories are distributed nontrivially within PLM weights (Vinod et al., 2026). A recent perspective by Hunklinger & Ferruz (2025) places these approaches within the broader emerging landscape of explainable protein AI.

These methods primarily ask which concepts, features, or interactions can be decoded from a representation. Our question is complementary: rather than assigning semantic labels to individual latent features, we ask how native three-dimensional organization is expressed in the geometry of the representation itself across network depth.

A.2 Layer-Wise Organization of Biological Knowledge

Several studies show that biological information is not distributed uniformly across PLM depth. Layer-wise probing has long suggested that intermediate representations can outperform final representations for particular structural or functional properties (Belinkov, 2021; Jawahar et al., 2019; Pinto, 2025). Recent PLM-specific studies make this depth dependence more explicit. Siji et al. (2026) analyze multiple PLM families, scales, and pretraining objectives using probing and perturbation measurements and find that a substantial fraction of downstream-relevant computation is concentrated before the final layers, framing the phenomenon in terms of depth inefficiency. Independently, Whitfield et al. (2026) perform high-resolution layer-wise probing across ESM2 and AMPLIFY models and observe an ordering in which physicochemical properties and local motifs are most accessible earlier, followed by secondary-structure information and later domain-level concepts.

These findings motivate studying intermediate representations, but the underlying question differs from ours. Probe accuracy or contribution to an output does not directly specify whether a layer preserves the native physical organization of residues. We therefore treat layer depth as an axis along which physically anchored structural organization can be measured directly, rather than inferring geometry from the decodability of downstream labels.

A.3 Representation Geometry, Topology, and Native Structure

A separate line of work studies hidden representations as geometric objects. In language and vision models, intrinsic dimensionality, anisotropy, neighborhood structure, and topological summaries reveal substantial reorganization across depth (Ansuini et al., 2019; Valeriani et al., 2023; Birdal et al., 2021; Naitzat et al., 2020; Williams et al., 2021). Similar ideas have recently been applied to protein representations. Valeriani et al. (2023) analyze hidden-state geometry in PLMs, while Rissom et al. (2025) compare molecular embedding spaces through both global pairwise relations and local neighborhoods, demonstrating that embedding-space geometry can expose how biological information is organized without task-specific model training.

The most directly related work is Beshkov & Malthe-Sørenssen (2025), which represents proteins and their latent representations with Square-Root Velocity shape analysis and graph filtrations. Their study establishes that PLM representation geometry is strongly depth-dependent, reports nonlinear evolution of effective shape-space dimensionality in ESM2 and Ankh, and finds that structural fidelity is strongest near but before the final layer. It further shows that structural preservation depends on neighborhood scale, with strong local and intermediate-scale features. This work provides an important foundation for studying PLM representations as geometric objects rather than only through supervised probes.

Our study builds on this geometric perspective but addresses a broader set of questions. First, we separate a global population-level description of whole-protein shape variability from a local, physically anchored measure

of within-protein native-neighborhood fidelity. Second, we evaluate these quantities across substantially different PLM architectures, model scales, general-purpose language models, and protein-oriented checkpoints rather than restricting the analysis to a small set of PLM families. Third, secondary-structure conditioning directly resolves which residue contexts contribute to the multi-scale local-fidelity profile. Finally, we evaluate whether local native-neighborhood fidelity has operational value for representation selection rather than treating the geometric profile only as a descriptive observation.

Structure-aware and multimodal protein models provide a complementary setting in which structural information is supplied explicitly. SaProt incorporates structure-derived tokens into the protein vocabulary (Su et al., 2023), while ESM3 jointly models sequence, structure, and function (Hayes et al., 2025). Recent layer-wise analysis of ESM3 shows that explicitly supplied physical modalities undergo depth-dependent fusion within its residual stream (Steenwyk, 2026). Our present study instead emphasizes sequence-only PLMs when asking where native structural alignment appears without direct structural input. Extending the same analyses to structure-aware models would address the related but distinct question of how explicitly supplied geometry is reorganized across depth.

A.4 Layer Selection and Depth-Dependent Representation Utility

The practical use of intermediate representations has been studied through supervised layer-wise probing, learned layer mixtures, and task-specific representation search (Belinkov, 2021; Jawahar et al., 2019; Skean et al., 2025; Pinto, 2025). These approaches establish that the final layer need not be optimal, but typically require downstream labels to determine which layer should be used. Recent depth analyses of PLMs similarly demonstrate that downstream-relevant computation can concentrate in intermediate layers (Siji et al., 2026), but do not provide a native-structure-based criterion for selecting a layer independently of the downstream task.

Unsupervised representation-comparison methods provide another route. Similarity measures such as CKA and related geometric comparisons characterize correspondences between representation spaces (Kornblith et al., 2019; Williams et al., 2021), while local embedding-space analyses can quantify neighborhood agreement without training an additional predictor (Rissom et al., 2025). Such criteria, however, need not be tied to the physical structure of the specific protein.

LTF differs in that its selection signal is externally anchored to native three-dimensional residue neighborhoods. A structural reference set is used to identify the layer whose latent neighborhoods most consistently preserve physical neighbors across structural scales, before any downstream labels are observed. The resulting criterion is therefore not a claim that intermediate layers are generally superior, but a protein-specific way to select representations when preservation of native structural organization is relevant.

A.5 Positioning of the Present Work

Taken together, prior work establishes four complementary facts: biologically meaningful concepts can be decoded from PLMs; such concepts vary across depth; PLM representation spaces undergo nontrivial geometric reorganization; and intermediate representations can be useful downstream. Our contribution lies in connecting these observations through a common physical reference.

Specifically, GMP/IMD provides a descriptive global view of how whole-protein representation-shape variability changes across layers, while LTF measures the preservation of native residue neighborhoods within individual proteins. Their comparison reveals that global geometric compression and local native-structure fidelity need not evolve together. The cross-model analysis then asks which aspects of this organization recur across architectures, scales, and training regimes; secondary-structure conditioning resolves the residue contexts associated with the local multi-scale signal; and downstream experiments test whether this physically anchored signal can guide model-adaptive, downstream-label-free layer selection. This positioning deliberately avoids treating intermediate-layer usefulness itself as novel, and it does not assume explicit coordinate encoding or a universal geometric trajectory.

B GMP AND IMD TECHNICAL DETAILS

This section specifies the shape-analysis steps underlying the descriptive global analysis in Section 2.2.

B.1 DISCRETE-TO-CONTINUOUS INTERPOLATION

Interpolation provides the curve derivatives required by the SRV transform. Given an ordered sequence of physical Cα coordinates or latent residue representations, we fit quadratic spline segments with continuous values and first derivatives at the knots. Quadratic splines provide the lowest-order smooth interpolation needed by the SRV construction while avoiding unnecessary higher-order curvature between observed residue positions. Curves are parameterized on $[0, 1]$ using arc length and discretized at $n = 1{,}200$ points. This fixed sampling grid allows proteins of different sequence lengths to be analyzed under the same numerical resolution.

Only residue-aligned points are included: prompt, boundary, and other non-residue tokens are removed before interpolation following Appendix D.1.

B.2 SQUARE-ROOT VELOCITY REPRESENTATION

The SRV transform expresses a curve through its local velocity (Srivastava et al., 2010). For $\mathbf{c} : [0, 1] \to \mathbb{R}^d$,

$$\mathbf{q}(t) = \begin{cases} \dot{\mathbf{c}}(t)/\sqrt{\|\dot{\mathbf{c}}(t)\|_2}, & \|\dot{\mathbf{c}}(t)\|_2 \neq 0, \\ \mathbf{0}, & \|\dot{\mathbf{c}}(t)\|_2 = 0. \end{cases}$$

The resulting representation lies in $L^2([0, 1], \mathbb{R}^d)$. Differentiation removes translation. Parameterization and scale normalization are separate operations and should not be conflated with the SRV transform.

B.3 SHAPE-SPACE ALIGNMENT

Rotational alignment removes the arbitrary orientation of curves within each ambient space. For fixed dimension d, the rotational equivalence class is

$$[q] = \{Rq : R \in SO(d)\},$$

and pairwise alignment solves

$$\hat{R} = \arg \min_{R \in SO(d)} \|q_1 - Rq_2\|_{L^2},$$

using SVD-based Procrustes alignment.

We do not additionally optimize over curve reparameterizations. Residue order provides a common biological parameterization of the physical backbone and its corresponding latent trajectory, so this order is preserved throughout the analysis.

Physical curves ($d = 3$) and latent curves ($d = d_{\text{hidden}}$) are analyzed in their respective shape spaces; they are not embedded into a common cross-dimensional tangent space.

B.4 KARCHER MEAN AND TANGENT-SPACE ANALYSIS

Tangent-space analysis summarizes between-protein variation around a representative shape within each curve family. Writing $d_{\mathcal{S}}$ for the shape-space distance, the Karcher mean is

$$\boldsymbol{\mu} = \operatorname*{argmin}_{p \in \mathcal{S}} \sum_{i=1}^{N} d_{\mathcal{S}}^2(p, \mathbf{q}_i).$$

The iterative procedure maps shapes to the current tangent space, averages the tangent vectors, updates the center through the exponential map, and repeats to convergence, typically within 5–10 iterations. Each shape is then represented by $\mathbf{v}_i = \operatorname{Log}_{\boldsymbol{\mu}}(\mathbf{q}_i) \in T_{\boldsymbol{\mu}}\mathcal{S}$. Tangent PCA describes variation within this family. A separate mean and tangent construction are used for each latent layer and for the physical reference.

B.5 INTRINSIC MANIFOLD DIMENSIONALITY

IMD summarizes the concentration of the tangent covariance spectrum. Using the same convention as Section 2.2,

$$C^{(\ell)} = \frac{1}{N-1} \sum_{i=1}^{N} \mathbf{v}_i^{(\ell)} \mathbf{v}_i^{(\ell)\top}, \qquad \text{IMD}^{(\ell)} = \frac{(\sum_k \lambda_k^{(\ell)})^2}{\sum_k (\lambda_k^{(\ell)})^2}.$$

This is the participation ratio of the nonzero spectrum: it equals one when variance is concentrated in one direction and p for equal variance in p directions. The physical reference, $\text{IMD}_{\text{phys}} \approx 3.88$, describes variability across protein shapes, not the three coordinates of one residue. Similar IMD values in different ambient dimensions provide comparable spectral summaries without establishing native-neighborhood preservation.

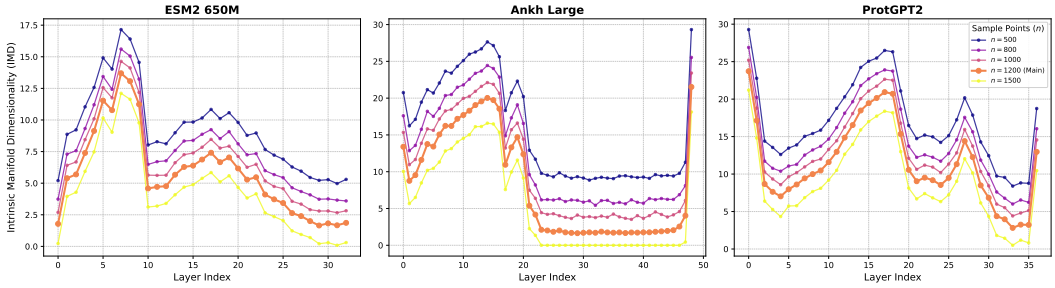

Figure 4: Sampling-density ablation. Sampling density affects absolute numerical IMD values while preserving the qualitative layer-wise trajectory across the tested resolutions within each model. The main experiments use $n = 1{,}200$.

B.6 Sampling Resolution

Sampling resolution tests the sensitivity of the discretized SRV analysis. Figure 4 compares the tested resolutions; $n = 1{,}200$ is the default used in the main experiments.

The comparison supports qualitative stability across the tested sampling resolutions, without asserting identical numerical values, convergence at a particular resolution, or a universal trajectory across model families.

C Dataset and Model Details

Dataset roles distinguish structural reference selection from downstream evaluation and delimit the scope of each result.

C.1 SCOPe Structural Reference Set

SCOPe provides the default physical reference for GMP and LTF. The reference is derived from SCOPe 2.08, with missing or corrupted PDB files removed during preprocessing.

Composition. The final dataset consists of 1,098 protein domains spanning seven structural classes: All Alpha (198), All Beta (195), Alpha/Beta (200), Alpha+Beta (189), Membrane & Cell Surface (108), Small Proteins (142), and Designed Proteins (66). This composition provides broad coverage across distinct protein structural regimes while maintaining a balanced reference for the layer-wise geometric analyses.

The same SCOPe collection is also used for fold classification, making that evaluation in-domain with respect to the structural reference. CATH contact prediction and ECPred classification are cross-dataset evaluations.

C.2 CATH 4.2 Contact and Distance Prediction

CATH evaluates whether the selected representations support prediction on a separate structural dataset. We use 500 single-domain proteins from CATH-S40, with sequence identity reduced below 40% and lengths of 50–300 residues. The 350/75/75 train/validation/test split is stratified by CATH topology. Ground-truth Cα distances are discretized into 20 bins; the probe and distance-MAE calculation are given in Appendix D.4. Long-range contact evaluation uses residue pairs with sequence separation > 24, distinct from the > 6 separation used for distance MAE.

C.3 ECPred Enzyme Classification

ECPred tests representation utility at different levels of enzyme annotation. The filtered subset contains 800 enzymes from UniProtKB/Swiss-Prot, with homology reduction at 40% sequence identity and lengths below 400 residues. Structures are obtained through UniProt-to-PDB mapping or high-confidence AlphaFold2 predictions (pLDDT > 70). Proteins without either source are excluded. The train/validation/test counts are 560/120/120, with EC-class stratification. The L1 classes in this subset are oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases.

Tables 5 and 6 report label cardinality and class composition. The fine-grained L4 task uses 187 labels within this subset; the full ontology size is not the classification target.

Table 5: EC label cardinality for the filtered 800-enzyme subset. Chance accuracy denotes uniform prediction over labels in the subset.

EC level	Unique labels	Chance accuracy
Level 1	6	16.7%
Level 2	38	2.6%
Level 3	97	1.0%
Level 4	187	0.5%

Table 6: EC-L1 class counts in the filtered ECPred subset.

EC L1 Class	Train	Val	Test	Total	Fraction
1. Oxidoreductases	98	21	21	140	17.5%
2. Transferases	140	30	30	200	25.0%
3. Hydrolases	123	26	27	176	22.0%
4. Lyases	84	18	18	120	15.0%
5. Isomerases	56	12	12	80	10.0%
6. Ligases	59	13	12	84	10.5%
Total	560	120	120	800	100%

These distributions describe the composition of the evaluation subset; they do not establish a mechanism linking functional categories to geometry.

C.4 ProteinGym Exploratory Mutation Evaluation

ProteinGym extends the evaluation to mutation-effect scoring. The subset contains 15 assays from ProteinGym v1.2 (217 substitution assays), selected for an experimentally determined structure at resolution ≤ 2.5 Å, at

least 500 measured single-site variants, protein length ≤ 500, and coverage of stability, binding, enzymatic, and organismal-fitness assays. The assay list, zero-shot scoring protocol, and results are in Appendix G. This exploratory subset is distinct from the four evaluations used in the main aggregate ranking.

C.5 DeepLoc Applicability Analysis

DeepLoc supplies a task-dependence check beyond the primary structure-sensitive evaluations. We use the curated SaProtHub DeepLoc split and test macro-F1 over three seeds on ESM2-650M, Ankh Base, and CARP-640M. LTF selection uses the SCOPe structural reference without DeepLoc labels; the supervised oracle uses DeepLoc validation performance. Full results appear in Appendix F.12.

C.6 Models and Analysis Coverage

Model specifications identify the frozen backbones used in downstream evaluation. Table 7 lists the ten dedicated PLMs in that comparison; all use publicly available pretrained weights without fine-tuning.

Table 7: The ten dedicated PLMs used in the main downstream comparison. Enc/Dec: encoder/decoder; MLM: masked language modeling; CLM: causal language modeling; FIM: fill-in-the-middle.

Model	Architecture	Params	Layers	Hidden Dim	Training Obj.
ESM2-35M Lin et al. (2023)	Transformer (Enc)	35M	12	480	MLM
ESM2-150M Lin et al. (2023)	Transformer (Enc)	150M	30	640	MLM
ESM2-650M Lin et al. (2023)	Transformer (Enc)	650M	33	1280	MLM
Ankh Base Elnaggar et al. (2023)	Transformer (Enc-Dec)	450M	48	768	Span MLM
Ankh Large Elnaggar et al. (2023)	Transformer (Enc-Dec)	1.5B	48	1536	Span MLM
ProtGPT2 Ferruz et al. (2022)	Transformer (Dec)	738M	36	1280	CLM
CARP-38M Yang et al. (2024)	CNN (ByteNet)	38M	16	1024	MLM
CARP-76M Yang et al. (2024)	CNN (ByteNet)	76M	33	1024	MLM
CARP-640M Yang et al. (2024)	CNN (ByteNet)	640M	56	1280	MLM
ProtMamba Sgarbossa et al. (2025)	SSM (Mamba)	107M	16	1024	CLM+FIM

The descriptive analysis additionally includes general-purpose Qwen2.5-7B and Llama-3.1-8B Base/Instruct models and protein-oriented InstructBioMol checkpoints. The precise GMP and LTF panels differ and are enumerated in Figures 2 and 3, respectively. The ten-model downstream table should not be read as the coverage of every descriptive figure. Layer selection uses hidden layers indexed from 1 to L, with the final hidden layer at L; the fixed-depth rule uses $\lfloor pL \rfloor$.

D EXPERIMENTAL PROTOCOLS

These protocols separate structural layer selection from supervised probing and specify the shared evaluation settings.

D.1 REPRESENTATION EXTRACTION

Aligned residue representations provide a common basis for local analysis and downstream probes. Each backbone is frozen, and per-residue hidden states are extracted from every analyzed layer. Contact prediction and LTF retain the $n_p \times d_{\text{hidden}}$ residue tensor for protein p; fold and EC classification use mean pooling over residues. Representations are cached by model, layer, and protein identifier in HDF5 files.

For general-purpose LLMs, amino acids are space-separated for residue-level extraction; prompt, boundary, and other non-residue tokens are excluded. InstructBioMol uses its native protein-sequence convention. Physical coordinates are external references and are not supplied as model inputs. Residue identities must remain aligned between coordinates and hidden states.

D.2 LTF COMPUTATION AND MEAN-OVER-SCALE SELECTION

The LTF protocol measures preservation of native neighborhoods before any downstream task labels are used. For each layer and $k \in \{1, \ldots, 19\}$, construct directed k-NN graphs without self-edges. Physical distances are Euclidean distances between Cα coordinates; latent distances are cosine distances. The default null samples points from an isotropic Gaussian matched to the layer's empirical mean and variance. Ten null samples per protein are averaged. The Frobenius-norm differences are then averaged over reference proteins as in Eq. 3.

The default selector is the mean over neighborhood scales:

$$\mathcal{S}(\ell) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} S^{(\ell)}_{\text{LTF}}(k), \qquad \ell^* = \underset{\ell \in \{1,\ldots,L\}}{\arg\max}\ \mathcal{S}(\ell), \quad \mathcal{K} = \{1, \ldots, 19\}.$$

This is exactly the selection rule in Eq. 4. The max-over-scale variant is a separate aggregation ablation (Appendix F.2.3). The integer k counts neighbors; it is neither a sequence-separation threshold nor a physical-distance cutoff.

D.3 SECONDARY-STRUCTURE-CONDITIONED LTF

Secondary-structure conditioning localizes the LTF signal to neighborhoods centered on different residue contexts. DSSP assignments are grouped into helix (H/G/I), strand (E/B), and coil (T/S/C/$-$). Importantly, the full protein is processed unchanged: physical, latent, and null k-NN graphs are first constructed over all aligned residues, and conditioning is applied only when aggregating over anchor residues.

For residue r of protein i, the latent neighborhood mismatch at layer ℓ and scale k is

$$d_i^{(\ell,k)}(r) = \left\| A^{(k)}_{\text{phys},i}[r,:] - A^{(k)}_{\text{lat},i,\ell}[r,:] \right\|_2^2. \tag{5}$$

Because both rows are binary k-NN adjacency vectors with k_{eff} valid neighbors,

$$d_i^{(\ell,k)}(r) = 2\left(k_{\text{eff}} - \left|\mathcal{N}^{(k)}_{\text{phys},i}(r) \cap \mathcal{N}^{(\ell,k)}_{\text{lat},i}(r)\right|\right), \tag{6}$$

so the row-wise mismatch is determined directly by the number of native physical neighbors retained in latent space. Under the default construction, $k_{\text{eff}} = k$ for valid anchors.

Let $\mathcal{R}_{i,c}$ denote the set of anchor residues in protein i assigned to secondary-structure class $c \in \{\text{helix}, \text{strand}, \text{coil}\}$. The corresponding class-conditioned discrepancy is obtained by restricting the row-wise contribution to anchors in $\mathcal{R}_{i,c}$, while candidate neighbors remain drawn from all valid residues in the protein:

$$D^{(\ell,k)}_{\text{lat},i,c} = \left[\sum_{r \in \mathcal{R}_{i,c}} d_i^{(\ell,k)}(r) \right]^{1/2}. \tag{7}$$

The null discrepancy $D^{(k)}_{\text{null},i,c}$ is defined analogously by replacing $A^{(k)}_{\text{lat},i,\ell}$ with the null k-NN graph. The secondary-structure-conditioned LTF score is then

$$S^{(\ell)}_{\text{LTF},c}(k) = \frac{\mathbb{E}_i\left[D^{(k)}_{\text{null},i,c}\right]}{\mathbb{E}_i\left[D^{(\ell,k)}_{\text{lat},i,c}\right]}. \tag{8}$$

Thus, Helix-LTF, Strand-LTF, and Coil-LTF quantify the fidelity of neighborhoods *centered* on residues of the corresponding class; they do not construct separate graphs within isolated secondary-structure fragments. In particular, a helix-centered anchor may have strand or coil residues among its physical or latent neighbors.

D.4 Probing Heads and Task Metrics

Lightweight probes evaluate the utility of the selected frozen representations under a shared task-specific architecture.

Contact prediction. A 2-layer 1D CNN refines local sequence contexts: Layer 1: 64 filters, kernel size k=5, stride 1, padding 2, ReLU activation. Layer 2: 64 filters, same configuration. Pairwise features z_{ij} are constructed by broadcasting and concatenating refined features h_i and h_j for all residue pairs. A linear projection maps to 20 distance bins, followed by symmetrization $M_{\text{sym}} = 0.5(M + M^\top)$ and softmax. Loss: multi-class cross-entropy over distance bins.

MAE computation from distance bins. The 20 distance bins span 0–4Å, 4–8Å, …, 72–76Å, >76Å, with bin centers $c_b \in \{2, 6, 10, \ldots, 74, 78\}$Å. For each residue pair (i, j), the probing head outputs a softmax probability distribution $\{P_b\}_{b=1}^{20}$. The predicted distance is computed as the **expected value** under this distribution: $\hat{d}_{ij} = \sum_{b=1}^{20} P_b \cdot c_b$. The Mean Absolute Error is then: $\text{MAE} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} |\hat{d}_{ij} - d_{ij}^{\text{true}}|$, where $\mathcal{P}$ is the set of all residue pairs with sequence separation > 6. The pairwise prediction matrix is symmetrized before computing MAE: $M_{\text{sym}} = 0.5 \times (M + M^\top)$.

Fold classification. A linear classifier (logistic regression with L2 regularization) operates on mean-pooled representations. The classifier has no hidden layers, keeping the probe architecture fixed when comparing representations.

EC classification. A hierarchical MLP with shared trunk: Input layer ($d_{\text{hidden}} \to 256$) $\to$ two shared hidden layers ($256 \to 256$, ReLU) $\to$ four branching linear heads (one per EC level). Loss: weighted sum of per-level cross-entropy, $\mathcal{L} = \sum_{l=1}^{4} 1.5^{l-1} \mathcal{L}_{\text{CE}}^{(l)}$, with inverse-frequency class weighting.

Mutation effect prediction. No probing head (zero-shot). Masked marginal log-likelihood ratios are computed directly from the PLM's language modeling head fed with layer-ℓ representations. See Appendix G for details.

D.5 Training Hyperparameters

Shared hyperparameters control for probe configuration when comparing layer selectors. Table 8 gives the task-specific settings.

Table 8: Shared probe hyperparameters. ES denotes early stopping and pat. denotes patience in epochs.

Task	Optimizer	LR	Epochs	Batch	Regularization
Contact Pred.	Adam	1e-3	10	16	Dropout 0.1; ES pat. 3
Fold Classif.	Adam	1e-3	50	32	L2 weight decay 1e-4
EC Classif.	Adam	1e-3	15	32	Dropout 0.2; ES pat. 4

The probing seeds are $\{42, 2024, 7\}$; uncertainty summaries, where given, reflect variability across these seeds. Hyperparameters were selected on the ESM2-650M validation set and then fixed across models and selectors. Searched learning rates were $10^{-4}, 10^{-3}, 10^{-2}$; EC hidden widths were 128, 256, 512; contact-CNN filter counts were 32, 64, 128. The model-specific tuning diagnostic is in Appendix F.10. Random layer baselines additionally use the ten draws specified below. Null sampling is distinct from deterministic reuse of cached representations.

D.6 Definitions of the Ten Comparison Methods

Explicit method definitions distinguish downstream-label-free selection from supervised layer optimization in the ten-method comparison.

Final layer. Use hidden layer L before the language-model head.

Random (Uniform). Draw uniformly from $\{1, \ldots, L\}$. Ten independent layer draws use seeds $\{0, \ldots, 9\}$; results are averaged over draws.

Random (Top-50%). Use the same protocol, restricted to $\{\lceil L/2 \rceil, \ldots, L\}$.

Fixed depth (80%). Select $\lfloor 0.8L \rfloor$. The expanded sweep tests 60–90% in five-percentage-point increments (Appendix F.1). Each proportion is a model-independent rule.

Mean pool (all layers). Average hidden states across layers, $H_{\text{pool}} = L^{-1} \sum_{\ell=1}^{L} H^{(\ell)}$, before probing.

IMD-selected. Select $\arg\min_\ell |\mathrm{IMD}^{(\ell)} - \mathrm{IMD}_{\mathrm{phys}}|$, with physical reference approximately 3.88, using the same SCOPe reference set. This is a descriptive global-shape statistic used here as a selection baseline.

Pairwise rank correlation (PRC). For each protein, compute Spearman correlation between physical Euclidean and latent cosine distances over all unique residue pairs. Average those correlations over reference proteins:

$$S_{\mathrm{PRC}}(\ell) = \frac{1}{|\mathcal{D}_{\mathrm{ref}}|} \sum_{p \in \mathcal{D}_{\mathrm{ref}}} \rho_{\mathrm{Spearman}}\big(\mathrm{vec}_{i<j} D_p^{\mathrm{phys}}, \mathrm{vec}_{i<j} D_p^{\mathrm{latent},\ell}\big).$$

Choose the layer maximizing this score. Separation-filtered and Euclidean variants are reported in Appendix F.9.

LTF-selected. Apply the mean-over-scale rule of Eq. 4 on the SCOPe structural reference. The selector requires native structures but no downstream labels, gradients, or parameter updates. It scores local neighbor identity across $k = 1, \ldots, 19$.

Learned weighted sum. Jointly train softmax layer weights and the probe using downstream labels: $H_{\mathrm{mix}} = \sum_\ell \frac{\exp \alpha_\ell}{\sum_j \exp \alpha_j} H^{(\ell)}$. Initialize $\alpha_\ell = 0$ and use the probe learning rate.

Grid-search oracle. Train the probe at every candidate layer and select using the task's validation metric (minimizing MAE or maximizing an accuracy/F1 metric as appropriate). Test performance is then reported at the selected layer. This is a supervised reference, not a label-free selector.

D.7 Statistical Testing Protocols

The aggregate statistical analysis compares ten methods across ten PLMs and four evaluations: Contact, Fold, EC-L1, and EC-L4 ($N = 40$ model–task blocks). For the hierarchical EC task, we use the two endpoint levels—EC-L1 and EC-L4—to represent coarse- and fine-grained functional prediction, respectively. EC-L2 and EC-L3 are reported separately in Appendix E.2, but are not added as independent blocks to the aggregate ranking in order to avoid overweighting a single hierarchical task family relative to Contact and Fold. Friedman testing gives $p < 0.001$ with $k = 10$ methods and $df = 9$; the Nemenyi critical difference is 2.14 at $\alpha = 0.05$. Removing the in-domain fold task leaves three cross-dataset evaluations ($N = 30$), with the same $k = 10$ and $df = 9$, Friedman $p < 0.001$, and Nemenyi critical difference 2.47. Statistical tests are computed from full-precision model-level scores; per-model values are rounded for presentation. The aggregate ranks and groups are reported in Appendices E.3 and E.4.

Mutation prediction and DeepLoc are outside this comparison. The extended 16-selector ablation in Appendix F.2 uses a separate rank pool and is distinct from the ten-method Friedman/Nemenyi analysis.

D.8 Computational Budget

The computational budget distinguishes representation extraction, structural scoring, and downstream probe training. Table 9 summarizes resource estimates; CPU and GPU hours refer to different hardware and should be read separately.

Table 9: Computational costs. Hardware is identified in each row; the total combines the listed CPU and GPU hour estimates.

Block	Reported hours	Notes
Representation extraction (10 PLMs, all layers)	40–60 h (A100)	One-time; cached to disk
LTF computation (all layers)	20–30 h (CPU)	Parallelizable per model
Grid Search Oracle (all layers × 4 tasks × 10 models)	80–120 h (A100)	Most expensive block
Probing head training (10 methods × 4 tasks × 10 models)	80–100 h (A100)	3 seeds per run
Total	**220–310 h**	

E Downstream Results and Statistical Analyses

Task-specific performance and aggregate rankings characterize the utility of layer selection across models. The aggregate comparison covers ten methods; the contact matrix gives model-level results for six of them.

E.1 Contact Prediction across Ten PLMs

The contact matrix gives model-level context for the main aggregate MAE result. Table 10 compares six methods. Model-wise fixed-depth results are reported in Appendix F.1.

Table 10: Contact-distance MAE (Å, ↓) for six methods across ten dedicated PLMs. Bold marks the LTF row. Model-wise fixed-depth results are given in Table 14.

	ESM2-35M	*ESM2-150M*	*ESM2-650M*	*Ankh Base*	*Ankh Large*	*ProtGPT2*	*CARP-38M*	*CARP-76M*	*CARP-640M*	*ProtMamba*	*Mean*
Final Layer	9.8	8.9	7.97	8.6	8.2	9.4	10.1	9.5	8.8	9.7	9.10
Fixed Depth (80%)	9.2	8.1	7.45	8.3	7.6	9.2	9.9	9.2	8.5	9.5	8.70
IMD-Sel.	9.6	8.6	8.75	8.4	8.1	9.3	10.0	9.4	8.7	9.6	9.04
PRC-Sel.	10.2	9.4	9.50	9.1	8.8	9.8	10.5	10.3	9.3	10.1	9.70
LTF (Ours)	**9.2**	**8.1**	**7.29**	**8.0**	**7.6**	**9.0**	**9.6**	**8.9**	**8.2**	**9.3**	**8.52**
Oracle	9.0	7.9	7.29	7.8	7.4	8.8	9.4	8.7	8.0	9.1	8.33

E.2 EC Accuracy and Macro-F1

Accuracy and macro-F1 provide complementary summaries of the imbalanced EC classification task. Table 11 reports ESM2-650M results at all four EC levels under the shared probing protocol. The ordering between LTF and final representations is task-level dependent; this comparison does not identify a causal separation of functional and structural features.

Table 11: ESM2-650M EC accuracy and macro-F1 (↑). Bold identifies the best value among the displayed downstream-label-free methods at each level.

	Accuracy				**Macro-F1**			
Method	L1	L2	L3	L4	L1	L2	L3	L4
Final Layer	**.77**	.61	.53	.47	**.72**	.48	.38	.35
Fixed Depth	.72	.58	.52	.46	.66	.45	.37	.34
IMD-Selected	.73	.59	.51	.43	.67	.46	.36	.31
LTF-Selected	.76	**.63**	**.55**	**.49**	.70	**.51**	**.41**	**.38**
Grid Search Oracle	.77	.64	.56	.49	.72	.52	.42	.38

E.3 Aggregate Statistical Comparison

The aggregate comparison in Table 12 evaluates ten methods across ten PLMs and 4 evaluations: Contact, Fold, EC-L1, and EC-L4 ($N = 40$). LTF has mean rank 2.73 and is the only downstream-label-free method in the top Nemenyi group, alongside the supervised learned weighted sum and grid-search oracle. Friedman testing gives $p < 0.001$; the Nemenyi critical difference is 2.14. Statistical tests are computed from full-precision model-level scores; per-model values are rounded for presentation. The extended 16-selector ranks in Appendix F.2 concern a separate comparison pool.

Table 12: Mean ranks (↓) and Nemenyi groups for the ten-method comparison ($k = 10$, $N = 40$). Friedman $p < 0.001$ ($df = 9$); Nemenyi CD $= 2.14$ ($\alpha = 0.05$). Main Table 2 summarizes the same comparison.

Method	Mean Rank	Nemenyi Group	Note
Grid Search Oracle	1.23	A	Performance ceiling
Learned Weighted Sum	2.12	A	Supervised mixture
LTF-Selected (Ours)	**2.73**	**A**	**Label-free selector**
Fixed Depth (80%)	5.72	B	Heuristic; partially effective
IMD-Selected	5.83	B	Geometric alternative
Mean Pool (All Layers)	6.23	BC	No-selection strategy
Final Layer	6.73	BC	Naive default
PRC-Selected	7.36	BCD	Distance correlation diagnostic
Random (Top-50%)	8.01	CD	Stronger lower bound
Random (Uniform)	9.04	D	Lower bound

E.4 Aggregate Excluding In-Domain Fold Classification

Removing fold classification distinguishes the cross-dataset evaluations from the in-domain SCOPe task. Table 13 summarizes the comparison on Contact, EC-L1, and EC-L4 ($N = 30$) alongside the full $N = 40$ summary. Both use the same ten-method pool; neither includes DeepLoc, mutation scoring, or the additional neighborhood selectors.

Table 13: Aggregate summaries with and without the in-domain fold task for the ten-method pool. Mean ranks are ↓; Δ is the $N = 30$ rank minus the $N = 40$ rank.

	Full (N=40)		**Cross-dataset** (N=30)		
Method	Rank	Group	Rank	Group	Δ
Grid Search Oracle†	1.23	A	1.18	A	−0.05
Learned Wt. Sum†	2.12	A	2.13	A	+0.01
LTF-Selected	**2.73**	**A**	**2.69**	**A**	**−0.04**
Fixed Depth (80%)	5.72	B	5.80	B	+0.08
IMD-Selected	5.83	B	5.82	B	−0.01
Mean Pool	6.23	BC	6.21	B	−0.02
Final Layer	6.73	BC	6.71	BC	−0.02
PRC-Selected	7.36	BCD	7.31	BC	−0.05
Random (Top-50%)	8.01	CD	8.01	BC	0.00
Random (Uniform)	9.04	D	9.14	C	+0.10
Cross-dataset: Friedman $p < 0.001$, CD $= 2.47$ ($k = 10$, $df = 9$, $N = 30$)					

Removing fold classification leaves the ranking highly stable: LTF changes only from 2.73 to 2.69 and remains the only downstream-label-free method in the top Nemenyi group alongside the two supervised layer-optimization references.

F LAYER-SELECTION ROBUSTNESS AND DIAGNOSTIC CONTROLS

These analyses characterize model adaptivity, sensitivity to analysis choices, native-reference specificity, and the task-dependent scope of LTF selection. Unless stated otherwise, structural scores use the SCOPe reference and contact MAE is evaluated on CATH.

F.1 FIXED-DEPTH SWEEP

The fixed-depth sweep tests how a common depth fraction compares with a model-specific structural selector. For $p \in \{0.60, 0.65, \ldots, 0.90\}$, select $\lfloor pL \rfloor$ and use the shared contact probe. Regret is the model-wise MAE difference from the supervised oracle; worst regret is the maximum across the ten PLMs. Table 14 reports results for three representative models together with ten-model aggregate summaries.

Table 14: Fixed-depth contact results: MAE (Å, ↓), with selected layers in brackets. Model depths are 33, 16, and 56, respectively. Mean and worst regret cover ten PLMs. †Supervised oracle.

Selector	ESM2-650M	CARP-38M	CARP-640M	Mean MAE	Worst regret
Fixed 60%	8.62 [19]	10.18 [9]	8.73 [33]	9.02	1.34
Fixed 65%	8.68 [21]	10.09 [10]	8.46 [36]	8.92	1.39
Fixed 70%	8.75 [23]	9.98 [11]	8.24 [39]	8.78	1.49
Fixed 75%	7.40 [24]	9.87 [12]	8.36 [42]	8.61	0.59
Fixed 80%	7.45 [26]	9.87 [12]	8.50 [44]	8.70	0.54
Fixed 85%	7.64 [28]	9.79 [13]	8.66 [47]	8.92	0.77
Fixed 90%	7.84 [29]	9.71 [14]	8.77 [50]	9.06	1.00
LTF-selected	7.29 [25]	9.60 [15]	8.20 [40]	8.52	0.20
Oracle†	7.29	9.40	8.00	8.33	0.00

The strongest fixed rule by mean contact MAE is 75% depth (8.61 Å), close to the 8.52 Å obtained by LTF. The practical distinction is clearer in model adaptivity and reliability: the lowest observed worst-case regret among the tested fixed rules is 0.54 Å, compared with 0.20 Å for LTF. The representative columns illustrate the model dependence of a useful depth fraction: 70% gives the best tested fixed result for CARP-640M, while the 60–90% range does not reach CARP-38M's selected layer 15. The selected-depth range across the ten PLMs is 68.8–93.8%; the representative depths are 75.8% for ESM2-650M, 93.8% for CARP-38M, and 71.4% for CARP-640M. This comparison supports model-adaptive selection without claiming a large average gain over every fixed rule.

F.2 SIMPLE NEIGHBORHOOD SELECTORS AND LTF COMPONENTS

Simple selectors test which aspects of native-neighborhood fidelity support representation selection. Structural selectors use the SCOPe reference and frozen representations without downstream labels. The component comparison separates physical anchoring, multi-scale aggregation, and null normalization. Mean ranks are computed over the extended pool of 16 selectors; Table 15 compares ten members of this pool. This ablation is distinct from the ten-method Friedman/Nemenyi comparison.

Table 15: Selector design and performance. Contact MAE (Å, ↓) is averaged over ten PLMs; EC-L4 accuracy is ↑. Mean ranks (↓) belong to the extended 16-selector pool. †Supervised reference.

Selector	Physical anchor	Multi-scale	Null norm.	MAE	EC-L4	Rank
Final Layer	No	No	No	9.10	0.455	7.1
Fixed Depth (80%)	No	No	No	8.70	0.462	5.4
PRC-Selected	Global	No	No	9.70	0.398	8.7
kNN-overlap@8	Local	No	No	8.62	0.466	4.5
Trustworthiness@8	Local	No	No	8.84	0.451	6.0
Continuity@8	Local	No	No	8.77	0.456	5.7
LTF@8	Local	No	Yes	8.60	0.468	4.2
LTF without null	Local	Yes	No	8.56	0.474	3.6
Full LTF	Local	Yes	Yes	8.52	0.479	3.1
Grid Search Oracle†	–	–	–	8.33	0.488	1.4

The dominant contribution comes from physically anchored local neighborhood preservation. Even the simple k-NN overlap criterion at $k = 8$ achieves a mean contact MAE of 8.62 Å, substantially stronger than global pairwise-distance correlation. Adding null normalization (LTF@8), multi-scale aggregation without the null, and finally the full LTF construction progressively improves mean MAE from 8.62 to 8.60, 8.56, and 8.52 Å, respectively.

These results indicate that the main signal is the preservation of native local neighbor identity, while multi-scale aggregation and null normalization provide complementary refinement and robustness. The full LTF formulation therefore consolidates a strong local structural signal rather than deriving its performance from a single specialized component.

F.2.1 Representative Selected Layers

Selected-layer comparisons illustrate model-adaptive localization of the structural region. Table 16 separates these indices from the component-level performance summaries.

Table 16: Representative selected layers on the SCOPe reference. Overlap, trustworthiness, and continuity use $k = 8$. The contact oracle uses CATH validation labels.

Model	L	Overlap	Trust.	Cont.	No null	Full LTF	Oracle
ESM2-650M	33	24	22	23	25	25	25
Ankh Base	48	37	34	36	38	39	39
CARP-640M	56	39	36	38	41	40	41

Full LTF selects the contact-oracle layer for ESM2-650M and Ankh Base and an adjacent layer for CARP-640M (25/39/40 versus 25/39/41). Simpler local metrics also locate this structural region, generally at somewhat shallower layers. The comparison does not imply a universal optimal depth.

F.2.2 Reliability and Worst-Case Regret

Worst-case regret measures architecture-dependent failure relative to the supervised single-layer oracle:

$$R_{\max} = \max_{m} \big[\mathrm{MAE}_{\mathrm{selector},m} - \mathrm{MAE}_{\mathrm{oracle},m}\big].$$

Worst-case regrets are computed from full-precision model-level MAEs; the contact matrix presents rounded per-model values.

Table 17: Worst-case contact regret across ten PLMs (Å, ↓), computed from full-precision model-level MAEs. †Supervised reference.

Selector	Worst regret
Final Layer	1.12
Fixed Depth (80%)	0.54
PRC-Selected	2.21
kNN-overlap@8	0.46
Trustworthiness@8	0.91
Continuity@8	0.76
LTF@8	0.42
LTF without null	0.32
Full LTF	0.20
Grid Search Oracle†	0.00

Full LTF reduces observed worst regret from 0.54 to 0.20 Å relative to Fixed Depth (80%), and from 0.46 to 0.20 Å relative to k-NN overlap@8. PRC has worst regret 2.21 Å, illustrating that a global use of physical distances does not itself ensure reliable layer selection. Full LTF provides a robust multi-scale implementation of local physical-neighborhood fidelity while remaining downstream-label-free. Worst regret is a descriptive reliability measure and does not establish statistical superiority.

F.2.3 Mean versus Maximum over Scales

This ablation separates the default mean aggregation from a max-over-scale alternative. The two scores are $\mathcal{S}_{\mathrm{mean}}(\ell) = |\mathcal{K}|^{-1} \sum_{k\in\mathcal{K}} S^{(\ell)}_{\mathrm{LTF}}(k)$ and $\mathcal{S}_{\max}(\ell) = \max_{k\in\mathcal{K}} S^{(\ell)}_{\mathrm{LTF}}(k)$, with $\mathcal{K} = \{1, \ldots, 19\}$. The mean is the default throughout this paper.

The two aggregation rules select layers within two positions of each other across all models. Both rules therefore identify nearby structural regions; mean aggregation remains the default for layer selection.

Table 18: Mean (default) versus maximum over neighborhood scales. Selected layers and contact P@L (↑) are shown for both aggregation rules.

	mean		max		
Model	ℓ^*	P@L	ℓ^*	P@L	$\lvert\Delta\ell\rvert$
ESM2-35M	9	.35	9	.35	0
ESM2-150M	24	.43	23	.42	1
ESM2-650M	25	.51	24	.48	1
Ankh Base	39	.46	38	.45	1
Ankh Large	38	.52	37	.51	1
ProtGPT2	29	.36	27	.35	2
CARP-38M	15	.30	14	.29	1
CARP-76M	23	.36	22	.35	1
CARP-640M	40	.44	38	.43	2
ProtMamba	11	.34	11	.34	0
Mean $\lvert\Delta\ell\rvert$			1.0		
Agreement ($\lvert\Delta\ell\rvert \leq 2$)			10/10 (100%)		

F.3 Selected Structural Region

The three highest mean-LTF layers characterize the selected structural region on ESM2-650M. Table 19 summarizes their probe performance; the final-layer reference has P@$L/2 = 0.66$ and MAE 7.97 Å.

Table 19: Top-three mean-LTF-ranked layers for ESM2-650M, with mean LTF scores and contact/EC probe performance.

Rank	**Layer**	$\mathcal{S}(\ell)$	**P@L**	**P@$L/2$**	**MAE (Å)**	**EC L4**
1	25	3.12	0.51	0.68	7.29	0.49
2	24	3.08	0.48	0.69	7.40	0.47
3	26	2.95	0.50	0.67	7.45	0.46

F.4 Reference-Set Robustness

Reference-set analyses examine whether a reusable structural collection can support selection across changes in size, composition, and source dataset.

F.4.1 Reference-Set Size

Subset-size sensitivity assesses the amount of structural reference data required for stable selection. We use SCOPe subsets of $N \in \{50, 100, 200, 500, 1098\}$ and ten bootstrap resamples per size. The cross-model summary averages selected-layer standard deviations over ten models. Table 20 also reports ESM2-650M selections and MAE.

Table 20: Reference-set size sensitivity. ESM2-650M layer and contact MAE are shown with the cross-model standard deviation of selected layers over ten bootstrap resamples.

N	ℓ^*	MAE (Å)	Cross-model $\sigma(\ell^*)$	**Note**
50	24	7.40	∼1.8	Moderate noise
100	25	7.29	∼0.9	Near-converged
200	25	7.29	∼0.4	Converged
500	25	7.29	∼0.2	Stable
1098	25	7.29	∼0.1	Full set (reference)

The ESM2 selection is layer 25 from $N = 100$ onward. This supports estimating a useful region from a smaller diverse reference set in the tested setting, without implying that one sample-size threshold applies to every model or structural distribution.

F.4.2 Structural-Class Composition

Class-restricted references test sensitivity to the structural composition of SCOPe. LTF is recomputed within each of the seven classes, with layer shifts measured relative to the full-set selection. Tables 21 and 22 summarize the ESM2 results and the cross-model selection summaries.

Table 21: ESM2-650M selections under class-restricted SCOPe references ($L = 33$). Δ is the absolute layer deviation from the full-set selection; contact MAE is in Å.

SCOPe Subset	N	ℓ^*	Δ	MAE (Å)	**Note**
Full Set (reference)	1098	25	–	7.29	Default
All Alpha	198	25	0	7.29	Exact agreement
All Beta	195	24	1	7.40	Adjacent
Alpha/Beta	200	25	0	7.29	Exact agreement
Alpha+Beta	189	26	1	7.45	Adjacent
Small Proteins	142	24	1	7.40	Adjacent
Membrane	108	25	0	7.29	Exact agreement
Designed	66	23	2	8.75	Slight deviation (small N)

Table 22: Mean and maximum absolute layer deviations across seven SCOPe classes for four representative PLMs.

	ESM2-650M	**Ankh Large**	**CARP-76M**	**ProtMamba**
$\|\Delta\|_{\text{mean}}$	0.7	0.9	1.0	1.1
$\|\Delta\|_{\max}$	2	2	3	2
Agreement ($\|\Delta\| \leq 2$)	7/7	7/7	6/7	7/7

Across the seven class-restricted references, the selected layers remain close to the full-reference solution. Mean deviations range from only 0.7 to 1.1 layers across the four representative PLMs, and 27 of the 28 model–class comparisons fall within two layers of the full-reference selection. These results show that LTF identifies a stable structural region despite substantial changes in the structural-class composition of the reference set, while retaining modest model- and class-specific variation.

F.4.3 Transfer across Structural Reference Datasets

Changing the source dataset tests whether the structural selection can be reused without target-task labels. LTF is recomputed independently on SCOPe, CATH 4.2, the ECPred structural subset, and the ProteinGym structural subset; compare the full layer-wise profiles and the five highest-ranked layers to SCOPe.

Main Table 4 reports the compact ESM2-650M result: layers 25, 24, 26, and 25, respectively. Cross-reference profile correlations are 0.76, 0.72, and 0.71 for CATH, ECPred, and ProteinGym, with four of five top-ranked layers retained in each comparison. The performance changes are below 0.5%, 0.3%, and 0.2%, respectively. The SCOPe reference row uses "–" for profile correlation and 5/5 for its reference top-five set. These are reference-transfer comparisons; the 0.81 correlation in Appendix F.11 instead concerns LTF and perturbation sensitivity.

F.5 Sequence-Locality Controls

Sequence-local controls test whether selection is dominated by covalent and nearby sequence neighbors. Candidate pairs are masked *before* constructing the physical, latent, and null neighborhoods, so all three graphs use the same allowed edges and each valid anchor retains k neighbors. The strongest mask removes candidates with $|i - j| \leq 24$.

Table 23: LTF selection after removing sequence-local candidates ($|i - j| > 24$). ρ is the Spearman correlation with the unmasked layer-wise profile. Results for three representative models and ten-model means are shown.

Model	L	ℓ^* **original**	ℓ^* **masked**	$\|\Delta\ell\|$	ρ
ESM2-650M	33	25	24	1	0.65
CARP-640M	56	40	38	2	0.61
ProtMamba	16	11	10	1	0.59
Mean over ten PLMs	–	–	–	1.4	0.63

Although removing sequence-local candidates changes the detailed LTF profiles, the selected structural region remains stable: ESM2-650M shifts from layer 25 to 24, CARP-640M from 40 to 38, and ProtMamba from 11 to 10, with a mean shift of 1.4 layers across the ten PLMs. Thus, the pre-final high-fidelity region persists even after nearby sequence neighbors are excluded. At the same time, the resulting profile changes confirm that the smallest-neighborhood regime contains a substantial sequence-local component, whereas the layer-selection signal extends beyond this contribution.

F.6 NATIVE-STRUCTURE SPECIFICITY

Reference interventions test whether LTF is specific to native neighbor identity. Sequences, PLM representations, and latent graphs remain fixed; only the physical reference structure or graph changes. Control D1 rewires edges within sequence-separation bins while preserving each residue's out-degree and per-bin edge counts. Control D2 uses a length-matched non-native structure from a different SCOPe class. Both controls are averaged over ten seeds for the 1,098 reference proteins.

Table 24 reports native LTF, its difference from each control, and the native-normalized D1 difference. GSI measures sensitivity in the independent representation-perturbation analysis in Appendix F.11; it is not an outcome of changing the reference graph.

Table 24: Native-reference specificity on ESM2-650M. ΔLTF is native minus control; relative Δ is divided by native LTF. D1/D2 denote the reference interventions defined above. GSI comes from the separate representation-perturbation analysis (Appendix F.11).

Layer	Depth	Native LTF	ΔD1	Rel.ΔD1	ΔD2	GSI
6	18%	1.12	0.09	8.0%	0.14	0.93
17	52%	1.75	0.28	16.0%	0.40	1.15
23	70%	2.60	0.72	27.7%	0.91	1.53
25	76%	3.12	1.14	36.5%	1.38	1.86
26	79%	2.95	0.95	32.2%	1.15	1.69
33	100%	1.48	0.17	11.5%	0.25	1.18

Among the evaluated layers, native-reference specificity is strongest at the LTF-selected layer 25. Rewiring the physical neighborhood structure reduces the native advantage by 36.5% at layer 25, compared with only 11.5% at the final layer; the independent perturbation analysis shows the same concentration, with GSI increasing from 1.18 at the final layer to 1.86 at layer 25. Because the latent representations are held fixed while only the physical reference is changed, this intervention directly isolates sensitivity to native neighbor identity. Together, the two controls localize structural specificity to the same pre-final region identified by LTF.

F.7 NULL-MODEL SENSITIVITY

Null-model comparisons test sensitivity to the normalization reference. We compare isotropic Gaussian samples, within-protein shuffled latent representations, and cross-protein shuffled representations. Table 25 compares the selected layers.

Table 25: Selected layers under three null constructions. The comparison concerns selected-region robustness.

Null model	ESM2-650M	CARP-76M	ProtMamba
Isotropic Gaussian (default)	25	23	11
Within-protein shuffle	25	23	11
Cross-protein shuffle	24	22	11

The selected layers are unchanged when the isotropic Gaussian reference is replaced by within-protein shuffling, and shift by at most one layer under cross-protein shuffling. Thus, the high-fidelity region is stable across qualitatively different null constructions, indicating that the selected layer is not determined by a particular choice of normalization reference.

F.8 LATENT DISTANCE METRIC

Distance-metric sensitivity tests the choice of metric for latent k-NN graphs. We compare cosine and Euclidean distances while retaining physical Euclidean graphs, the isotropic Gaussian null, the SCOPe reference, and $k = 1, \ldots, 19$.

Cosine and Euclidean latent distances identify the same layer for five of the ten PLMs and an adjacent layer for the remaining five, giving a mean absolute shift of only 0.5 layers. The selected structural region is therefore highly stable to this basic choice of latent-space metric. Cosine distance is retained as the default throughout the main analysis.

Table 26: Selected layers with cosine and Euclidean latent distances.

Model	Cosine	Euclidean
ESM2-35M	9	9
ESM2-150M	24	24
ESM2-650M	25	25
Ankh Base	39	38
Ankh Large	38	37
ProtGPT2	29	29
CARP-38M	15	14
CARP-76M	23	22
CARP-640M	40	39
ProtMamba	11	11

F.9 Pairwise-Distance Diagnostic

The PRC diagnostic tests whether global agreement of pairwise distance rankings provides the same layer-selection signal as local preservation of native neighbor identity. The comparison reveals a systematic separation between these two notions of structural correspondence. Table 27 compares the selected layers; two controls examine pair filtering and latent distance.

Table 27: PRC and mean-LTF selected layers across ten PLMs; PRC scores use all residue pairs.

Model	L	ℓ^*_{PRC}	S_{PRC}	ℓ^*_{LTF}
ESM2-35M	12	3	0.41	9
ESM2-150M	30	5	0.46	24
ESM2-650M	33	6	0.49	25
Ankh Base	48	8	0.44	39
Ankh Large	48	9	0.47	38
ProtGPT2	36	7	0.38	29
CARP-38M	16	4	0.35	15
CARP-76M	33	6	0.39	23
CARP-640M	56	10	0.43	40
ProtMamba	16	4	0.36	11

F.9.1 Sequence-Separation Filtering

Filtering pairs tests the contribution of sequence-local distances to PRC. On ESM2-650M, PRC is recomputed using only pairs beyond each tested sequence separation (Table 28). This filters pairs in the global correlation, unlike the candidate masking before k-NN construction in Appendix F.5.

Table 28: PRC selections under sequence-separation filtering on ESM2-650M, with LTF as a reference. Contact MAE is in Å (↓).

Pair filter	Selected layer	S_{PRC}	MAE
All pairs	6	0.49	9.50
$\lvert i-j\rvert > 6$	8	0.42	9.35
$\lvert i-j\rvert > 12$	10	0.36	9.20
$\lvert i-j\rvert > 24$	12	0.28	9.10
LTF-selected (reference)	25	–	7.29

PRC shifts from layer 6 to layer 12 under the strongest filter, while the default LTF selection is layer 25. Removing increasingly local sequence pairs shifts PRC from layer 6 to layer 12, but it remains substantially earlier than the LTF-selected layer 25. The distinction between the two selectors therefore persists even after strong sequence-separation filtering: global pairwise-distance agreement and local native-neighborhood fidelity emphasize different regions of the representation hierarchy.

F.9.2 PRC Latent Distance

Changing the latent metric tests whether PRC's early-layer preference depends on cosine distance. Table 29 reports cosine and Euclidean results for three representative PLMs.

Changing the latent metric shifts the PRC-selected layer by only one position for each representative model and leaves downstream MAE qualitatively unchanged. The early-layer preference of PRC is therefore not an

Table 29: PRC selections using cosine and Euclidean latent distances; contact MAE is in Å (↓).

	ESM2-650M		CARP-76M		ProtMamba	
Latent Metric	ℓ^*	MAE	ℓ^*	MAE	ℓ^*	MAE
Cosine (default)	6	9.50	6	10.3	4	10.1
Euclidean	5	9.65	5	10.4	3	10.2

artifact of the cosine metric, further separating global pairwise-distance agreement from the pre-final native-neighborhood region identified by LTF.

F.10 Probe Hyperparameters across Architectures

Probe retuning checks sensitivity to the shared ESM2-derived hyperparameters. CARP-76M, Ankh Large, and ProtMamba are retuned over 27 contact and 18 EC configurations, varying learning rate, filter count or hidden width, and regularization. Table 30 reports the shared and model-tuned protocols and correlation of the ten-method rankings.

Table 30: Shared versus model-tuned probe settings. Contact MAE is in Å (↓); EC-L4 accuracy is ↑. Rank ρ compares the ten-method ordering between protocols.

		Shared		Model-Tuned			
Model	**Task**	LTF	Final	LTF	Final	**ΔLTF**	**Rank ρ**
CARP-76M	Contact MAE (Å)	8.9	9.5	8.7	9.3	−0.2	0.78
	EC L4 Acc.	.41	.39	.43	.41	+.02	0.77
Ankh Large	Contact MAE (Å)	7.6	8.2	7.5	8.1	−0.1	0.75
	EC L4 Acc.	.50	.48	.51	.49	+.01	0.72
ProtMamba	Contact MAE (Å)	9.3	9.7	9.1	9.5	−0.2	0.73
	EC L4 Acc.	.37	.35	.38	.36	+.01	0.72

Model-specific retuning modestly improves absolute probe performance while preserving substantial agreement in method ordering ($\rho = 0.72$–0.78). More importantly, the LTF-selected representation continues to outperform the final layer in all six model–task comparisons under the retuned protocols. These results show that the practical advantage of the selected region is not specific to the shared ESM2-derived probe configuration, while the shared protocol provides a controlled basis for the primary cross-model comparison.

F.11 Representation Perturbation Analysis

Representation perturbations test the sensitivity of frozen contact probes across selected depths. A contact probe is trained on clean representations at a layer and evaluated on perturbed representations from that same layer with the probe fixed. ESM2-650M layers 6, 17, 23, 25, 26, and 33 are tested. Isotropic noise is $\tilde{h}_i^{(\ell)} = h_i^{(\ell)} + \varepsilon \|h_i^{(\ell)}\|_2 z_i$, with $z_i \sim \mathcal{N}(0, I_d/d)$ and $\varepsilon \in \{0.02, 0.05, 0.10, 0.20, 0.50\}$. Local sequence-window shuffling permutes residue representations within windows of $w \in \{3, 5, 9, 15, 25\}$. It preserves the embedding distribution while disrupting the residue-to-representation assignment. Results are averaged over five perturbation realizations and three probe seeds.

Table 31: ESM2-650M contact MAE under isotropic perturbation. ΔMAE is perturbed minus clean MAE, in Å. Bold marks the largest degradation in each noise column.

		ε=0	ΔMAE				
Layer	**Depth**	**MAE**	0.02	0.05	0.10	0.20	0.50
6 (PRC)	18%	9.50	+0.04	+0.09	+0.20	+0.42	+1.05
17 (Mid)	52%	8.50	+0.06	+0.15	+0.32	+0.68	+1.55
23 (IMD)	70%	8.75	+0.08	+0.21	+0.45	+0.92	+1.90
25 (LTF)	**76%**	**7.29**	**+0.10**	**+0.26**	**+0.55**	**+1.12**	**+2.20**
26	79%	7.45	+0.09	+0.23	+0.48	+0.98	+2.00
33 (Final)	100%	7.97	+0.05	+0.12	+0.25	+0.52	+1.25
Ratio Δ(L25) / Δ(L33)			2.0	2.2	2.2	2.2	1.8

Table 31 reports absolute degradation, while Table 32 also normalizes degradation by the clean MAE (RDR). Perturbation sensitivity increases toward the high-LTF region and is largest at the selected layer 25 across all tested noise levels. After normalizing by clean prediction error, RDR remains strongly associated with LTF over all 33 layers ($\rho = 0.81$, $p < 0.001$). Thus, an independent perturbation-based measurement localizes representation sensitivity to the same pre-final region identified by native-neighborhood fidelity.

Table 32: Sensitivity at $\varepsilon = 0.10$. RDR is ΔMAE divided by clean MAE. Region names locate the sampled layers and do not denote universal phases.

Layer	Depth	S_{LTF}	MAE	ΔMAE	RDR	Region
6	18%	1.12	9.50	+0.20	2.1%	Early
17	52%	1.75	8.50	+0.32	3.8%	Intermediate
23	70%	2.60	8.75	+0.45	5.1%	LTF-adjacent
25	**76%**	**3.12**	**7.29**	**+0.55**	**7.5%**	**LTF-selected**
26	79%	2.95	7.45	+0.48	6.4%	LTF-adjacent
33	100%	1.48	7.97	+0.25	3.1%	Final
Spearman ρ (RDR vs. S_{LTF}, all 33 layers)			$\rho = 0.81,\ p < 0.001$			

F.11.1 Local Sequence-Window Shuffle

The shuffle comparison tests sensitivity to residue ordering at the selected and final layers. Table 33 reports sensitivity across sequence-window sizes. These perturbations affect sequence-local ordering and are distinct from changing the physical reference graph in Appendix F.6.

Table 33: Contact ΔMAE (Å) under local sequence-window shuffling. L25/L33 ratios compare degradation at the selected and final layers.

	ΔMAE at window size w				
Layer	w=3	w=5	w=9	w=15	w=25
L25 (LTF)	+0.35	+0.72	+1.30	+1.85	+2.40
L33 (Final)	+0.10	+0.22	+0.45	+0.70	+0.95
Ratio L25/L33	3.5	3.3	2.9	2.6	2.5

F.11.2 Displacement-Matched Sensitivity

Matching displacement provides a comparison between shuffle and isotropic perturbations at each layer. For the $w = 9$ shuffle, isotropic noise is calibrated per layer to match mean per-residue ℓ_2 displacement, with $\varepsilon_{\mathrm{matched}} \in [0.12, 0.18]$. The Geometric Sensitivity Index is $\mathrm{GSI} = \Delta\mathrm{MAE}_{\mathrm{shuffle}}/\Delta\mathrm{MAE}_{\mathrm{iso}}$.

Table 34: GSI under displacement-matched perturbations. The numerator and denominator are contact ΔMAEs in Å; GSI is their ratio.

Layer	Depth	ΔMAE$_{\mathrm{iso}}$	ΔMAE$_{\mathrm{shuffle}}$	GSI	Interpretation
6 (PRC)	18%	+0.30	+0.28	0.93	Early layer
17 (Mid)	52%	+0.48	+0.55	1.15	Lower ratio
23 (IMD)	70%	+0.62	+0.95	1.53	Intermediate ratio
25 (LTF)	**76%**	**+0.70**	**+1.30**	**1.86**	**LTF-selected layer**
26	79%	+0.65	+1.10	1.69	High
33 (Final)	100%	+0.38	+0.45	1.18	Lower ratio

Under displacement-matched perturbations, GSI peaks at the LTF-selected layer 25 (1.86), compared with 1.18 at the final layer. The selected representation is therefore disproportionately sensitive to disruption of residue organization relative to isotropic perturbation of comparable magnitude. Combined with the native-reference interventions in Appendix F.6, this provides two complementary signatures—one acting on the structural reference and one on the representation—that independently concentrate in the same pre-final region.

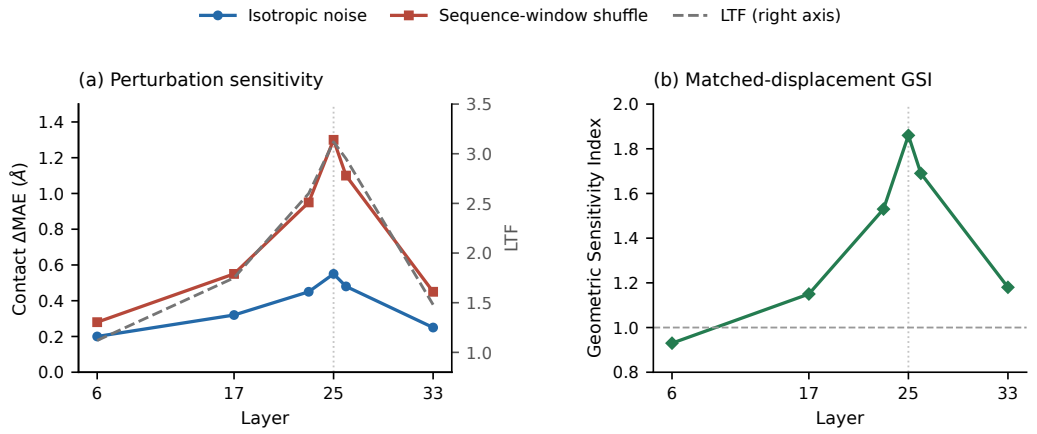

Figure 5: Perturbation sensitivity at six ESM2-650M layers. (a) Contact ΔMAE for isotropic noise ($\varepsilon = 0.10$) and sequence-window shuffle ($w = 9$), with LTF on the right axis. (b) Displacement-matched GSI. Points correspond to Tables 32 and 34; connecting segments are visual guides. The vertical line marks layer 25.

F.12 Task Dependence: DeepLoc

DeepLoc provides a complementary test of how the relationship between native-neighborhood fidelity and representation utility changes for a task that is less directly tied to residue-level structural geometry. Using the curated split described in Appendix C.5, we compare final, fixed-80%, LTF-selected, and supervised-oracle representations on ESM2-650M, Ankh Base, and CARP-640M. LTF selection uses the SCOPe structural reference without DeepLoc labels, whereas the oracle selects a layer using DeepLoc validation performance. Test macro-F1 is averaged over three seeds.

Table 35: DeepLoc test macro-F1 (↑). Selected layer indices precede the performance columns. Oracle is supervised; Δ is LTF minus Final.

Model	L	ℓ_{LTF}	$\ell_{80\%}$	ℓ_{oracle}	Final	Fixed	LTF	Oracle	Δ
ESM2-650M	33	25	26	31	0.684	0.662	0.654	0.692	-0.030
Ankh Base	48	39	38	47	0.711	0.694	0.688	0.717	-0.023
CARP-640M	56	40	44	53	0.615	0.607	0.595	0.624	-0.020
Mean	–	–	–	–	0.670	0.654	0.646	0.678	-0.024

For these three models, DeepLoc favors representations closer to the output of the network. Final-layer representations outperform the LTF-selected layers, and the supervised oracle consistently selects a near-final layer (31/33, 47/48, and 53/56). This behavior contrasts with the pre-final preference observed on the primary evaluations and demonstrates that the relationship between native structural fidelity and downstream representation utility is task dependent.

Rather than implying a universal preference for intermediate layers, these results sharpen the interpretation of LTF: it identifies layers with strong native-neighborhood organization, and such organization is particularly useful when it aligns with the information required by the target task. Determining how broadly this relationship extends across functional, evolutionary, and sequence-level prediction problems is an important direction for future study.

G MUTATION EFFECT PREDICTION: EXPLORATORY ANALYSIS

Mutation prediction explores the utility of selected representations beyond the primary contact, fold, and EC evaluations.

G.1 ZERO-SHOT SCORING PROTOCOL

Masked marginal scoring tests mutation effects without a trained downstream probe. For a single substitution at position i from residue a to b,

$$\Delta_\ell(a \to b \mid X) = \log p_\ell(b \mid X_{\setminus i}) - \log p_\ell(a \mid X_{\setminus i}),$$

where p_ℓ is obtained by applying the language-model head to the masked position's layer-ℓ representation. The protocol covers five PLMs: ESM2-35M/150M/650M and Ankh Base/Large. The oracle is a supervised layer-selection reference and must be distinguished from the label-free scoring/selection methods.

G.2 ASSAY SUBSET

Assay metadata establishes the scope and selection criteria of the exploratory panel. The 15 assays satisfy the structural availability, variant count, length, and category criteria in Appendix C.4. LTF is computed on these same proteins, so this is not the SCOPe-to-target transfer setting of the primary tasks. Structural availability may favor well-characterized proteins.

Table 36: Characteristics of the 15 ProteinGym assays. S: stability; B: binding; E: enzymatic activity; O: organismal fitness. The listed structures are experimental.

#	Protein	Organism	PDB	Res.(Å)	Len.	Variants	Cat.
1	avGFP	*A. victoria*	1GFL	1.90	238	2,573	S
2	GB1 domain	*S. pyogenes*	1PGA	2.07	56	1,045	S
3	TEM-1 β-lactamase	*E. coli*	1BTL	1.80	286	4,807	E
4	T4 lysozyme	*Phage T4*	2LZM	1.70	164	2,015	S
5	Ubiquitin	*H. sapiens*	1UBQ	1.80	76	1,370	B
6	PTEN	*H. sapiens*	1D5R	2.10	403	5,202	E
7	NUDT15	*H. sapiens*	5LPG	1.70	164	2,756	E
8	Src SH3 domain	*H. sapiens*	1SRL	2.00	64	1,082	B
9	WW domain (hYAP65)	*H. sapiens*	1PIN	1.50	34	608	S
10	λ repressor	*Phage* λ	1LMB	1.80	92	1,680	S
11	BRCA1 RING domain	*H. sapiens*	1JM7	1.85	110	1,756	B
12	MAPK1 / ERK2	*H. sapiens*	2ERK	2.30	360	6,320	E
13	AAV2 capsid VP1	*AAV-2*	1LP3	2.50	444	7,120	O
14	Spike RBD	*SARS-CoV-2*	6M0J	2.45	201	3,819	B
15	APH(3')-II	*E. coli*	1ND4	2.40	264	4,293	E
Category distribution: S:5, B:4, E:5, O:1						*Total: 46,446 variants*	

G.3 FIVE-MODEL RESULTS

The model-level summary reports average assay-wise Spearman correlation between masked marginal scores and measured fitness. Table 37 compares layer selectors on this panel. These results do not support extrapolation to the remaining five PLMs or to the full ProteinGym benchmark.

Table 37: Mutation-effect Spearman ρ (↑), averaged across the 15 assays for five PLMs. Bold marks LTF among the displayed downstream-label-free methods; the oracle is supervised.

	ESM2-35M	*ESM2-150M*	*ESM2-650M*	*Ankh Base*	*Ankh Large*	*Mean*
Final Layer	.35	.41	.44	.39	.43	.404
Fixed Depth (80%)	.36	.42	.45	.40	.44	.414
IMD-Selected	.35	.41	.43	.39	.43	.402
LTF-Selected (Ours)	**.37**	**.43**	**.46**	**.41**	**.45**	**.424**
Grid Search Oracle	.38	.44	.47	.42	.46	.434

The mean changes from 0.404 at the final layer to 0.424 with LTF selection. This exploratory difference is evaluated on the selected assay panel and should not be compared directly with another method's score on a different ProteinGym subset.

G.4 PER-ASSAY RESULTS

Per-assay reporting shows how the ESM2-650M summary is distributed over the panel. Table 38 summarizes the fifteen assay results and individual differences. Each assay contributes equally to the unweighted mean, irrespective of its variant count.

Table 38: ESM2-650M per-assay Spearman ρ (↑), with $\Delta\rho = \mathrm{LTF} - \mathrm{Final}$. W/T/L is 14/1/0.

#	Protein	Cat.	Final	LTF	$\Delta\rho$
1	avGFP	S	.42	.45	+.03
2	GB1 domain	S	.52	.54	+.02
3	TEM-1 β-lact.	E	.48	.51	+.03
4	T4 lysozyme	S	.49	.50	+.01
5	Ubiquitin	B	.39	.42	+.03
6	PTEN	E	.35	.37	+.02
7	NUDT15	E	.46	.48	+.02
8	Src SH3	B	.43	.44	+.01
9	WW domain	S	.38	.38	.00
10	λ repressor	S	.50	.52	+.02
11	BRCA1 RING	B	.33	.35	+.02
12	MAPK1 / ERK2	E	.47	.50	+.03
13	AAV2 capsid	O	.41	.43	+.02
14	Spike RBD	B	.52	.53	+.01
15	APH(3')-II	E	.44	.47	+.03

This exploratory assay panel does not establish a quantitative relationship between task-property dependence and structural fidelity.

H LIMITATIONS, FUTURE DIRECTIONS, AND BROADER IMPACT

The analyses in this work provide both a descriptive account of where native structural organization appears across model depth and a practical way to exploit this signal for representation selection. We summarize the implications of these findings, the scope of the current evidence, and directions for extending the analysis.

H.1 PRACTICAL IMPLICATIONS

A central practical consequence of our results is that native structural organization can provide a model-adaptive signal for choosing representations without downstream labels or exhaustive layer search. Across the evaluated PLMs, LTF consistently identifies pre-final representations with strong native-neighborhood fidelity and substantial utility on downstream prediction tasks. The strongest gains appear on structure-sensitive evaluations, where the selected layers improve contact prediction and fine-grained structural or functional tasks over the final representation.

Importantly, the selector is not tied to a single structural reference dataset. Reference-set transfer experiments show that LTF identifies similar layer-wise regions across SCOPe, CATH, ECPred, and ProteinGym subsets, while reference-size experiments indicate that stable selection can already emerge from relatively small structural collections. These results make LTF useful not only as an analysis tool, but also as a lightweight representation-selection criterion that can be reused across downstream settings.

H.2 LIMITATIONS

The present study characterizes native structural organization across a broad but finite set of models, datasets, and downstream evaluations. Several aspects therefore remain open.

Reference coverage. LTF requires residue-aligned structural references, and its robustness is established for the reference sets, structural classes, and subset sizes evaluated here. Although selection remains stable across multiple reference datasets and compositions, broader structural distributions may reveal additional variation.

Global versus local geometry. IMD and LTF intentionally capture different properties of a representation. IMD summarizes population-level variation in whole-protein representation shapes, whereas LTF measures local preservation of native residue neighborhoods. Their trajectories need not coincide, and global geometric compression alone should not be interpreted as native-neighborhood preservation.

Intervention scope. The native-reference controls and representation perturbations provide complementary evidence that structural specificity is concentrated near the LTF-selected region. These interventions establish sensitivity to native structural organization, but do not by themselves imply explicit coordinate encoding or a complete causal account of how geometric information is formed or used internally.

Evaluation coverage. The main downstream evaluation spans ten dedicated PLMs and multiple structure- and function-related tasks, while the broader general-purpose LLM and protein-oriented model analyses are primarily descriptive. Mutation-effect prediction and DeepLoc provide additional evidence beyond the main evaluation pool, but a substantially broader set of protein properties will be needed to determine how widely structural fidelity predicts downstream representation utility.

Statistical scope. The aggregate Friedman/Nemenyi analyses apply to the stated ten-method evaluation pools, while the extended neighborhood-selector study uses a separate comparison pool. These analyses should therefore be interpreted within their respective experimental settings rather than as universal rankings of representation-selection strategies.

H.3 FUTURE DIRECTIONS

Several directions naturally follow from the present findings. First, the observed relationship between native-neighborhood fidelity and downstream utility motivates a more systematic analysis of which protein or amino-acid properties depend primarily on local, mesoscopic, or global structural organization. Relating layer-wise LTF and GMP statistics to property-specific prediction performance could help identify which biological tasks benefit most from local versus global representation geometry.

Second, extending the analysis to explicitly structure-aware models such as SaProt and ESM3 would provide a useful contrast between structural organization that emerges from sequence-only pretraining and structure that is provided directly through model inputs or objectives. Such models could reveal whether explicit structural supervision produces uniformly high fidelity across depth or whether structure is still reorganized into distinct layer-wise regimes.

Third, broader interventions could test the functional role of these structural regions more directly. Layer-specific editing, controlled representation replacement, or targeted perturbation of selected structural scales could distinguish correlation from sufficiency and clarify how native-neighborhood information contributes to downstream predictions.

Finally, the reference-set results suggest the possibility of constructing compact, reusable structural calibration sets for representation selection. Understanding how to choose such reference proteins efficiently, and how selection transfers across increasingly diverse tasks and model families, is an important direction for practical use.

H.4 Broader Impact

Understanding where structural information is expressed in protein language models can make these representations easier to interpret and more efficient to use. A reusable structural reference set may reduce the need for repeated supervised layer search, lowering the computational and labeling cost of adapting pretrained protein models to new applications. More broadly, physically anchored representation analysis can provide a bridge between latent-space behavior and experimentally meaningful protein structure.

At the same time, the geometric signals studied here characterize model representations rather than replacing biological validation. High LTF or favorable downstream probe performance should not be interpreted as evidence that a model has learned a complete physical description of a protein, nor should computational predictions substitute for experimental confirmation in high-stakes biological applications.