When Does Structure Help? The Information Bonus of AlphaFold2 Representations over Protein Language Models

Kargi Chauhan [1]

Abstract

AI scientist systems increasingly choose biological foundation models before they choose experiments. In protein pipelines, this creates a concrete engineering and scientific question: *when is the cost of structural inference worth paying over a cheaper sequence-only model?* We introduce the **information bonus** (IB), a task-level metric that measures the linearly accessible advantage of frozen single-sequence AlphaFold2 Evoformer representations over frozen ESM-2 embeddings under protein-level cross-validation. Across binding affinity regression (PDBbind, n=5,680), conformational flexibility (ATLAS molecular dynamics, 268 proteins), and allosteric-site classification (AlloSigDB, n=9,925 residues), IB is sharply mechanism-dependent. ESM-2 dominates binding affinity (IB= -0.141; Pearson r=0.449 vs. 0.307) and binary flexibility (IB= -0.060; AUROC 0.824 vs. 0.764; p=0.0017). AF2 single representations give the only above-chance allostery predictions (IB= $+0.064$; AUROC 0.548 vs. 0.485), revealing long-range geometric signal not recovered from sequence alone. We also identify a residue-level leakage artifact: naive residue splits inflate RMSF performance by 27–39% depending on the representation, enough to reverse representation rankings. These results turn representation selection into a measurable decision for AI-for-science systems.

1. Introduction

Protein foundation models are now infrastructure for AI-assisted discovery. Sequence models such as ESM-2 can produce residue embeddings in seconds from a single amino-acid string (Lin et al., 2023). AlphaFold2 produces geometry-aware Evoformer representations, but at the price of structural inference, specialized software, and substantially higher compute (Jumper et al., 2021; Mirdita et al., 2022). For an autonomous or semi-autonomous AI scientist, this choice is not cosmetic. It determines latency, cost, experimental throughput, and sometimes the scientific conclusion itself (Lu et al., 2024).

The usual advice is that "structure matters for structural tasks." That is true but not operational. Binding affinity, flexibility, and allostery are all structural in some sense, yet they differ in the information that a model must recover. Binding sites are constrained by evolution. Flexible regions often leave sequence signatures of disorder, charge patterning, and low packing. Allostery, by contrast, depends on non-local communication through the folded contact network. A useful benchmark should distinguish these regimes rather than assume one representation is globally superior.

We propose the **information bonus** (IB): the held-out performance difference between the best AF2 representation and ESM-2 on the same task, using the same frozen linear probe and the same protein-level split. IB>0 means structural pre-training adds usable signal; IB<0 means sequence embeddings are sufficient or better. This framing makes representation choice measurable, cheap to estimate, and directly actionable inside AI-for-science workflows.

Contributions. We make four contributions. First, we define IB as a simple benchmark unit for deciding when structural representations are worth their cost. Second, we show that ESM-2 is stronger for binding affinity and binary flexibility, despite having no explicit coordinates. Third, we identify allostery as the clearest positive-IB regime, where AF2 is the only representation above chance. Fourth, we quantify a 27–39% residue-level cross-validation inflation effect in RMSF probing and show why protein-level GroupKFold is the correct protocol for per-residue tasks.

[1]University of California, Santa Cruz. Correspondence to: Kargi Chauhan <kchauha3@ucsc.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

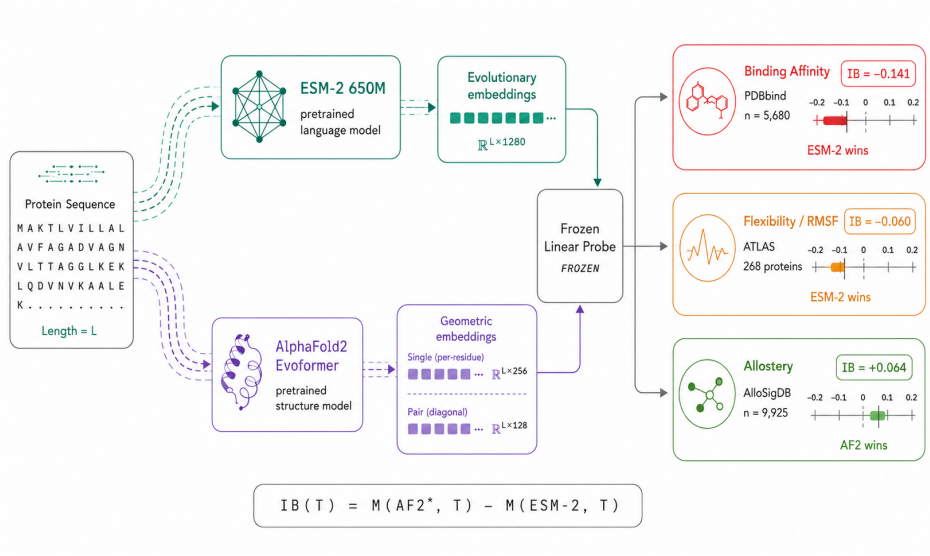

Figure 1. **Information Bonus evaluation framework.** Both models receive the same protein sequence. ESM-2 supplies sequence-only evolutionary embeddings; AlphaFold2 supplies Evoformer single and pair-diagonal representations. Frozen linear probes evaluate each representation on binding, flexibility, and allostery. The information bonus is the held-out performance difference between the best AF2 representation and ESM-2.

2. Related Work

Protein representation learning. AlphaFold2 uses Evoformer blocks to iteratively update residue-level and pairwise representations before coordinate prediction (Jumper et al., 2021). ESM-1b, ProtTrans, and ESM-2 instead learn from large protein sequence corpora and recover structural and functional regularities without coordinate supervision (Rives et al., 2021; Elnaggar et al., 2021; Lin et al., 2023). Recent work has shown that sequence embeddings alone can match or exceed structure-based methods on tasks such as virtual screening (Lam et al., 2024) and binding affinity prediction (Piao et al., 2025). Structure-aware language models such as SaProt combine sequence and structural tokens (Su et al., 2024), underscoring the open question studied here: which downstream tasks actually require explicit geometric information?

Probing and interpretability. Linear probes measure information that is already linearly accessible in a frozen representation (Alain & Bengio, 2016). TAPE established this style of evaluation for protein transfer learning (Rao et al., 2019), while later work showed that protein language models encode contacts, mutation effects, and functional constraints (Vig et al., 2021; Meier et al., 2021; Hie et al., 2022). Recent sparse autoencoder studies provide a complementary interpretability view: ESM-2 embeddings contain features aligned with domains, binding sites, structural motifs, Gene Ontology annotations, and protein families (Gujral et al., 2025; Simon & Zou, 2025). Our work asks a different question: when does AF2 add task-relevant information beyond those sequence features?

Benchmarks and leakage. PDBbind provides measured protein-ligand affinities (Liu et al., 2017); ATLAS provides molecular-dynamics-derived flexibility labels (Vander Meersche et al., 2024); and AlloSigMA/AlloSigDB organize experimentally supported allosteric mechanisms (Li et al., 2022). Molecular benchmarks are vulnerable to scaffold, homolog, and identity leakage (Yang et al., 2019; Varadi et al., 2022). Grouped cross-validation and careful split design are increasingly recognized as essential for protein tasks (Fang, 2023; Bushuiev et al., 2024; Bernett et al., 2024). We isolate a specific residue-level version of this problem: if residues from the same protein appear in both train and test folds, probes can exploit protein identity rather than transferable residue physics.

3. Information Bonus

Definition 1 (Information Bonus)**.** Let f_{AF2^*} be the stronger AF2 representation for task $\mathcal{T}$, chosen between single and pair-diagonal Evoformer features, and let f_{ESM} be ESM-2 650M embeddings. For scalar metric M under protein-level GroupKFold,

$$\mathrm{IB}(\mathcal{T}) = M(f_{\mathrm{AF2}^*}, \mathcal{T}) - M(f_{\mathrm{ESM}}, \mathcal{T}). \tag{1}$$

In practice, we evaluate both AF2 variants on every task and report all scores; f_{AF2^*} is selected post-hoc as the variant with the higher held-out metric. Because both variants are reported separately (Tables 2, 3), readers can verify the IB

for any AF2 configuration. This post-hoc selection gives structural representations a mild upward bias relative to the single ESM-2 baseline; we note that AF2 still yields negative IB on two of four evaluations despite this advantage.

IB>0 indicates usable structural signal; IB<0 indicates that sequence embeddings are sufficient or superior. We use Pearson r for regression and AUROC for classification. IB is intentionally a frozen-feature metric: it measures information available to a downstream system without expensive fine-tuning, model surgery, or task-specific representation learning.

4. Experimental Setup

Tasks. **Binding affinity** uses PDBbind (Liu et al., 2017) to predict $-\log_{10}(K_d)$ for 5,680 protein-ligand complexes from mean-pooled protein features alone (no ligand representation is provided to the probe; see Section 5 for interpretive consequences). **Flexibility** uses ATLAS (Vander Meersche et al., 2024), with 268 proteins and 50,426 residues, to predict RMSF by regression and by a balanced within-protein median classification task. **Allostery** uses AlloSigDB-derived labels (Li et al., 2022), with 47 proteins, 9,925 residues, and 4.8% positives, to identify allosteric sites.

Representations and probes. AF2 single features are final Evoformer single representations ($\mathbb{R}^{L\times 256}$). AF2 pair-diagonal features use z_{ii} ($\mathbb{R}^{L\times 128}$), a residue-level view of pairwise geometry. ESM-2 features are final-layer embeddings from `facebook/esm2_t33_650M_UR50D` ($\mathbb{R}^{L\times 1280}$). AF2 is run in single-sequence mode so both paths receive the same input. We use ridge regression (α=1.0) for scalar targets and class-balanced logistic regression (C=1.0) for binary targets, with ℓ_2-normalized frozen features and fixed seeds. Regularization strength is fixed across representations; because features are unit-normalized, probe capacity depends primarily on signal content rather than raw dimensionality.

Splits and statistics. The primary protocol is 5-fold protein-level GroupKFold: all residues from a protein are assigned to the same fold. This is essential for per-residue tasks because labels are correlated within proteins. We do not cluster proteins by sequence identity before splitting; homologous proteins may therefore appear in different folds, which could modestly inflate all scores but affects all representations equally. We report fold means, permutation baselines, paired fold tests, and Cohen's d where appropriate.

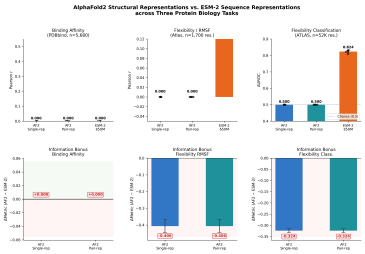

Figure 2. **Three-task overview.** ESM-2 wins binding affinity and binary flexibility, while AF2 is useful when the target depends on long-range geometry. Dots show held-out folds.

Table 1. **Main quantitative results.** Pearson r is used for regression; AUROC is used for classification.

Task	Best representation	Score	IB
Binding affinity	ESM-2	.449	-0.141
RMSF regression	AF2 pair diag.	.436	$+0.030$
Flexibility AUROC	ESM-2	.824	-0.060^{***}
Allostery AUROC	AF2 single	.548	$+0.064$

$^{***}p < 0.005$ against AF2 pair. IB is computed from unrounded fold-level scores; displayed values are rounded independently.

5. Results

5.1. Binding Affinity: Evolutionary Signal Wins

ESM-2 is the strongest representation for binding affinity, achieving Pearson r=0.449 compared with 0.307 for AF2 single and 0.278 for AF2 pair-diagonal features. The structural information bonus is therefore negative (IB$=-0.141$). This result is stable across folds.

An important caveat: because the probe receives only protein features (no ligand representation), the task effectively measures how well each representation captures protein-level correlates of affinity—pocket druggability, family-level binding propensity—rather than complex-specific complementarity. ESM-2's advantage is therefore best interpreted as stronger encoding of evolutionary and family-level binding constraints (Lam et al., 2024; Piao et al., 2025), not as evidence that sequence alone can predict ligand-specific K_d. The AF2 features describe the apo protein, not the ligand-bound complex, which further limits the structural path.

5.2. Flexibility: Static Geometry Helps Less Than Sequence Context

For continuous RMSF regression, AF2 pair-diagonal features are directionally best (r=0.436), slightly above ESM-2 (r=0.407). This weak positive IB is consistent with pair-diagonal geometry encoding burial, packing, and contact

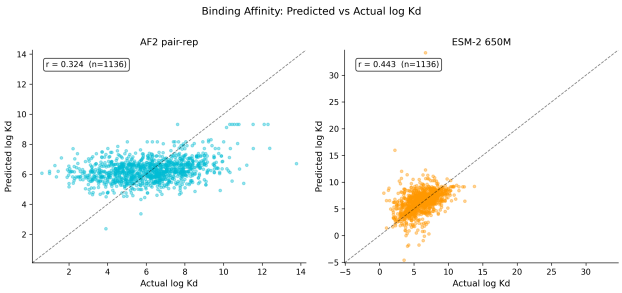

Figure 3. **Binding affinity scatter plot.** ESM-2 yields a higher predicted-versus-measured correlation ($r = 0.443$) than AF2 pair representations ($r = 0.324$) on the held-out test fold ($n = 1136$).

Table 2. **Flexibility results on ATLAS.**

Target	ESM-2	AF2 single	AF2 pair
RMSF r	.407	.401	**.436**
Median AUROC	**.824**	.762	.764
Permutation	.000/.500	.000/.500	.000/.500

density. However, the advantage is not statistically decisive with five folds.

For binary flexibility, ESM-2 wins cleanly in every fold, with AUROC 0.824 versus 0.764 for AF2 pair and 0.762 for AF2 single. The task asks for relative flexibility within a protein, and sequence models appear to capture the conserved signatures of disorder and mobility better than static AF2 features. This is a useful reminder: a task can be physically structural without requiring an explicitly structural representation.

5.3. Residue-Level Splits Inflate RMSF by 27–39%

Under naive residue-level 5-fold KFold, ESM-2 reaches r=0.672 on RMSF. Under protein-level GroupKFold, it drops to r=0.407, a 39.4% reduction. AF2 pair drops from approximately 0.600 to 0.436, a 27.3% reduction. Residue-level splitting therefore inflates apparent performance by 27–39% (relative reduction) depending on the representation, with the largest effect on ESM-2.

The mechanism is not subtle. Flexible proteins tend to contain many flexible residues; rigid proteins tend to contain many rigid residues. If residues from one protein are split across train and test sets, the probe can learn protein-level context instead of transferable residue-level physics. The error changes the scientific conclusion: ESM-2 appears dominant under leaky residue splits, while AF2 pair is directionally best under the correct protein-level RMSF regression protocol.

5.4. Allostery: Structure Provides the Missing Signal

Allostery is the only task with a clearly positive structural information bonus. AF2 single reaches AUROC 0.548, while

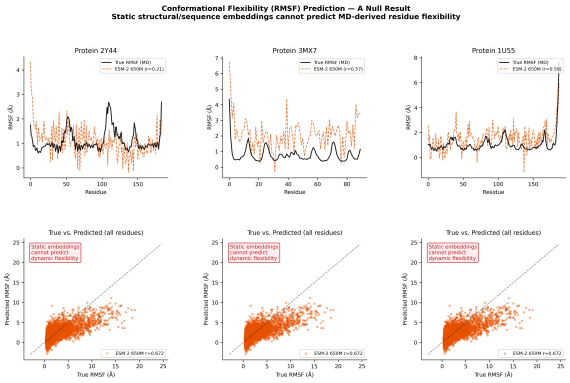

Figure 4. **RMSF detail and split sensitivity.** Residue-level splitting inflates performance by letting train and test residues share protein identity.

Table 3. **Allosteric-site classification** on AlloSigDB.

Representation	AUROC	IB
AF2 single	**.548**	+0.064
AF2 pair diag.	.497	+0.012
ESM-2	.485	—
Permutation	.500	—

AF2 single is the only representation above the chance baseline.

ESM-2 falls below chance at 0.485 and AF2 pair is near chance at 0.497. The absolute score is modest because the dataset is small (47 proteins, ∼9 per test fold) and imbalanced, so this result should be interpreted as directional evidence of structural signal rather than a statistically decisive finding. Only AF2 single contains above-chance signal.

This matches the biology. Allosteric residues are defined by their role in a three-dimensional communication network, not simply by local sequence patterns. The Evoformer repeatedly updates residue and pair representations through global structural context before coordinate prediction. That geometry appears to preserve signal about allosteric communication that ESM-2 does not linearly recover.

6. Analysis

The boundary is mechanistic, not architectural. The same AF2 representation is unnecessary for binding, weakly useful for continuous flexibility, and uniquely useful for allostery. This pattern is exactly what one would expect if sequence models absorb evolutionary regularities while structural models preserve information about spatial communication. IB therefore acts less like a leaderboard score and more like a diagnostic: it tells an AI scientist system what kind of biological signal a task requires.

Why sequence often wins. Evolution supplies supervision at a scale no structural database can match. Bind-

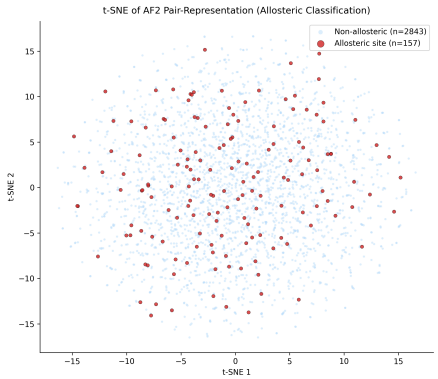

Figure 5. **AF2 representation space for allostery.** Allosteric residues show non-random clustering in AF2 single-representation space, consistent with geometric encoding of allosteric pathways.

ing constraints, disorder signatures, functional motifs, and family-level selection pressures recur across millions of sequences. ESM-2 can internalize these regularities without ever seeing explicit coordinates (Lin et al., 2023; Meier et al., 2021). When a downstream label is largely explained by those regularities, AF2's structural inference can add cost without adding linearly accessible information.

Why allostery is different. Allostery depends on residue placement inside a folded contact network. Sequence covariation can hint at coupling, but the task asks whether a particular residue participates in long-range communication. AF2 single features, produced after repeated Evoformer updates, carry enough global geometric context to rise above chance. This is the positive case for structural representations: not structure as decoration, but structure as the missing variable.

7. Discussion

Decision rule for AI scientist systems. Our results suggest starting with ESM-2 when the target is plausibly governed by evolutionary constraint, family-level function, disorder, or binding-site chemistry. We hypothesize that the AF2 cost is justified when the target explicitly depends on three-dimensional communication—as demonstrated here for allostery, and plausibly extending to protein-protein interfaces, cryptic pockets, and domain coupling (Cimermancic et al., 2016), though these extensions remain to be tested. When the mechanism is uncertain, estimating IB on a small labeled validation set before scaling structural inference provides a practical safeguard.

Benchmarking implication. The leakage result is as important as the representation result. Per-residue benchmarks must split by protein, not by residue, whenever labels are correlated within proteins. Otherwise, a model can appear to understand local biophysics while merely interpolating within proteins it has already seen. For AI-for-science systems that make downstream experimental decisions, this kind of evaluation error is not a formatting issue; it changes which model is selected.

Limitations. The allostery dataset is small and imbalanced (47 proteins, $\sim$9 per test fold), so AUROC should be interpreted as directional evidence of representation signal rather than a deployment-ready classifier; fold-level standard deviations and formal significance tests would strengthen this result. The RMSF comparison has limited fold-level statistical power. Single-sequence AF2 gives the cleanest matched-input comparison but may underestimate the advantage of full-MSA AF2 on geometry-dependent tasks; IB measured with full-MSA AF2 could differ substantially. Regularization hyperparameters are fixed across all representations rather than tuned per-representation via nested cross-validation, which could modestly affect absolute scores. Finally, linear probes measure accessible information; nonlinear fine-tuning could improve absolute performance, but would also change the cost and safety profile of the system.

8. Conclusion

We introduced the information bonus, a practical metric for deciding when single-sequence AlphaFold2 structural representations add usable information over ESM-2 sequence embeddings. Across three protein tasks, the answer is mechanism-dependent. ESM-2 wins when evolutionary signal is sufficient; in our experiments, AF2 provided a clear advantage only when long-range geometry was the missing information. The same study also shows that residue-level cross-validation can inflate RMSF performance by 27–39% and reverse representation rankings. For AI scientist systems, structure is powerful, but it should be treated as a measured scientific choice rather than a default assumption.

Broader Impact

This work improves the reliability of biological AI evaluation by making representation selection explicit and by identifying a common leakage mode in per-residue benchmarks. Better evaluation can reduce unnecessary structural inference costs and lower the risk of selecting the wrong model for closed-loop discovery. We do not anticipate direct dual-use concerns from the evaluation framework itself.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Bernett, J., Blumenthal, D. B., and List, M. Cracking the black box of deep sequence-based protein-protein interaction prediction. *Briefings in Bioinformatics*, 25(2), 2024. doi: 10.1093/bib/bbae076.

Bushuiev, A. et al. Revealing data leakage in protein interaction benchmarks. In *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2024.

Cimermancic, P., Weinkam, P., Rettenmaier, T. J., Bichmann, L., Keedy, D. A., Woldeyes, R. A., Schneidman-Duhovny, D., Demerdash, O. N., Mitchell, J. C., Wells, J. A., et al. CryptoSite: Expanding the druggable proteome by characterization and prediction of cryptic binding sites. *Journal of Molecular Biology*, 428(4):709–719, 2016.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., 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.

Fang, J. The role of data imbalance bias in the prediction of protein stability change upon mutation. *PLOS ONE*, 18 (3):e0283727, 2023. doi: 10.1371/journal.pone.0283727.

Gujral, O., Bafna, M., Alm, E., and Berger, B. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122(34):e2506316122, 2025. doi: 10.1073/pnas.2506316122.

Hie, B., Zhong, E. D., Berger, B., and Bryson, B. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Systems*, 13(4):274–285, 2022.

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

Lam, H. Y. I., Guan, J. S., Ong, X. E., Pincket, R., and Mu, Y. Protein language models are performant in structure-free virtual screening. *Briefings in Bioinformatics*, 25(6): bbae480, 2024. doi: 10.1093/bib/bbae480.

Li, L., Chang, S., Bhaskara, R. M., et al. AlloSigMA 2: Paving the way to designing allosteric effectors and to exploring allosteric effects of mutations. *Nucleic Acids Research*, 50(W1):W135–W144, 2022.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., and Wang, R. PDBbind: A comprehensive database for experimentally measured binding affinity data of biomolecular complexes. *Nucleic Acids Research*, 45(D1):D955–D963, 2017.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34, 2021.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. ColabFold: Making protein folding accessible to all. *Nature Methods*, 19(6):679–682, 2022.

Piao, H. et al. BindPred: A framework for predicting protein-protein binding affinity from language model embeddings. *bioRxiv*, 2025.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. S. Evaluating protein transfer learning with TAPE. *Advances in Neural Information Processing Systems*, 32, 2019.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., 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.

Simon, E. and Zou, J. InterPLM: Discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22:2107–2117, 2025. doi: 10.1038/s41592-025-02836-7.

Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X., and Yuan, F. SaProt: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations*, 2024.

Vander Meersche, Y., Cretin, G., Gonceareneo, A., Sterckx, Y. G. J., Gelly, J.-C., and de Brevern, A. G. ATLAS: Protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Research*, 52(D1): D384–D392, 2024.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. AlphaFold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 2022.

Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., and Rajani, N. F. BERTology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations*, 2021.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019.