PROTCOMPASS: BIOLOGICAL INFORMATION AND PREDICTIVE UTILITY IN PROTEIN REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein encoders are widely evaluated through structural probing, downstream prediction, and concept erasure, but these measurements can mix learned signal with information already present in the input or introduced by the evaluation itself. We introduce **ProtCompass**, a unified evaluation approach for separating these sources and identifying what protein representations genuinely contribute. ProtCompass audits what each encoder actually reads, compares structural probe scores with non-learned descriptions of the same inputs, evaluates downstream performance against simple sequence baselines, and controls concept-erasure experiments for the representation capacity removed together with the target property. Across diverse protein encoders and biological tasks, we find that structural probe scores can largely reflect retained input information, simple sequence statistics explain substantial downstream performance on several tasks, and many apparent task–property dependencies weaken once concept-erasure effects are properly controlled. These results provide a clearer way to determine what protein representations learn, what they add to biological prediction, and which properties downstream models actually use.

1 INTRODUCTION

Protein encoders map amino acid sequences, structures, or both into representations used in protein engineering, fitness prediction, function annotation, etc. (Nguyen & Hy, 2025; Wu et al., 2023). Those encoders include sequence language models such as ESM (Rives et al., 2019; Meier et al., 2021; Lin et al., 2023), ProtTrans (Elnaggar et al., 2022; Brandes et al., 2022), and Ankh (Elnaggar et al., 2023); structure-based models such as GearNet (Zhang et al., 2022) and PottsMPNN (Birnbaum & Keating, 2026; Hermosilla et al., 2020); multimodal models such as SaProt (Su et al., 2023) and ESM3 (Hayes et al., 2025; Wu et al., 2023; Zhang et al., 2023). Their representations are commonly assessed through downstream prediction, probing, and erasure. Downstream benchmarks measure how well an embedding supports biological prediction (Rao et al., 2019; Xu et al., 2022; Dallago et al., 2021; Notin et al., 2023); probes test whether structural or biochemical properties can be decoded from a frozen representation (Prabakaran & Bromberg, 2026); and erasure methods remove selected information from a representation and measure the resulting change in downstream performance (Elazar et al., 2021; Belrose et al., 2023). These evaluations are often interpreted as evidence about what information a representation encodes, which encoder is useful for a task, and what information a predictor depends on.

These measurements assume that the score mainly reflects the property being measured. In practice, however, part of the score comes directly from the input or intervention itself. Many structural probe labels are functions of the coordinates already given to structure encoders; for example, a backbone dihedral is fixed by four backbone atoms. Likewise, simple sequence features such as amino acid composition and length can already predict some downstream tasks. Erasing a property also removes representation capacity, so performance can drop even when the task does not specifically depend on that property. Thus, raw scores can mix learned information with signals built into the evaluation itself.

This leads to one question: *What biological information is accessible from protein representations, how do these representations support biological prediction, and what evidence links particular properties to predictive performance?* ProtCompass separates these effects by comparing each score

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.

with a simple control that measures what the input or intervention alone can explain (Figure 1). We first audit which inputs each encoder actually uses. For structural probing, we compare the encoder with simple descriptors computed directly from those same inputs and evaluated with the same probe. We call the resulting score the **floor**: the performance that can be achieved from the input alone. For downstream prediction, we compare pooled embeddings with a non-learned reference containing amino-acid composition and length. For erasure, we compare the property erasure with a null that removes the same share of representation variance while carrying little of the property, and then add the property back to test whether it selectively repairs its own erasure. These matched comparisons ask what the learned representation contributes beyond information or capacity already available to the measurement.

Across 33 encoders, 21 probes and ten rebuilt biological prediction tasks, the answer changes the interpretation of all three measurements. For structural probing, the decisive variable is whether the encoder's input already determines the label. On nine determined labels whose floors do not contain the answer directly, coordinate-reading encoders fall below a floor built from their own atoms on 186 of 200 encoder–label pairs, whereas sequence-only encoders fall below their sequence floor on none of 90. The same probe score therefore has different meanings across input modalities: for a coordinate encoder it largely measures how accessible a known input remains, while for a sequence-only encoder it reflects structural information inferred from sequence. Where the stored input does not determine the label, including solvent accessibility and crystallographic flexibility, some encoders exceed every floor we constructed. This separation persists on folds and labels that were frozen before scoring.

The same issue extends to downstream prediction and erasure. On downstream tasks, simple sequence statistics already explain a substantial part of performance, although large pretrained models add clear predictive value beyond them. For erasure, many apparent task–property dependencies weaken once we control for the amount of representation removed, and targeted follow-up tests leave only one robust candidate.

In conclusion, these results show that a score alone does not reveal what produced it; an appropriate control is needed to determine what information the representation actually adds. Structural probes should be compared with what the encoder's input already provides, downstream scores with simple sequence baselines, and erasure effects with controls that remove comparable representation capacity. ProtCompass brings these controls into a single benchmark to make each score easier to interpret.

2 Study Design

ProtCompass asks one question across probing, downstream prediction, and erasure: **what does a representation contribute beyond what the measurement already provides?** We therefore evaluate every score against a matched reference. Figure 1 summarizes the framework.

2.1 Dataset and Evaluation Setting

Probing data. We use CATH 4.3 (Sillitoe et al., 2020), containing 19,461 domains and 2,823,525 residues. Domains are split by CATH superfamily into 15,565 training, 1,948 validation, and 1,948 test proteins with zero group leakage; a topology-disjoint split is used for confirmation experiments. Residue- and pair-level probes are fitted on 3,000 training and 500 validation domains and scored on 1,000 test domains, while protein-level probes use the full split.

Evaluation protocol. All encoders remain frozen. Each probe is read with a linear model and an MLP-2 with two hidden layers of width 128. Model and floor readouts are selected on validation and evaluated once on test. Encoder–floor gaps receive 95% confidence intervals from 2,000 paired bootstrap samples of test proteins. Full probing, downstream, and erasure protocols are given in Appendices H.2, B.2, H.5, and D.2.

2.2 Protein Encoders and Input Audit

We evaluate 33 encoders: 9 sequence-based, 19 structure-based, and 5 multimodal models, spanning protein language models, geometric networks, inverse-folding models, and multimodal foundation models (Rives et al., 2019; Elnaggar et al., 2022; Lin et al., 2023; Elnaggar et al., 2023; Heinzinger et al., 2024; Schütt et al., 2017; Thomas et al., 2018; Jing et al., 2020; Gasteiger et al., 2020; Jumper

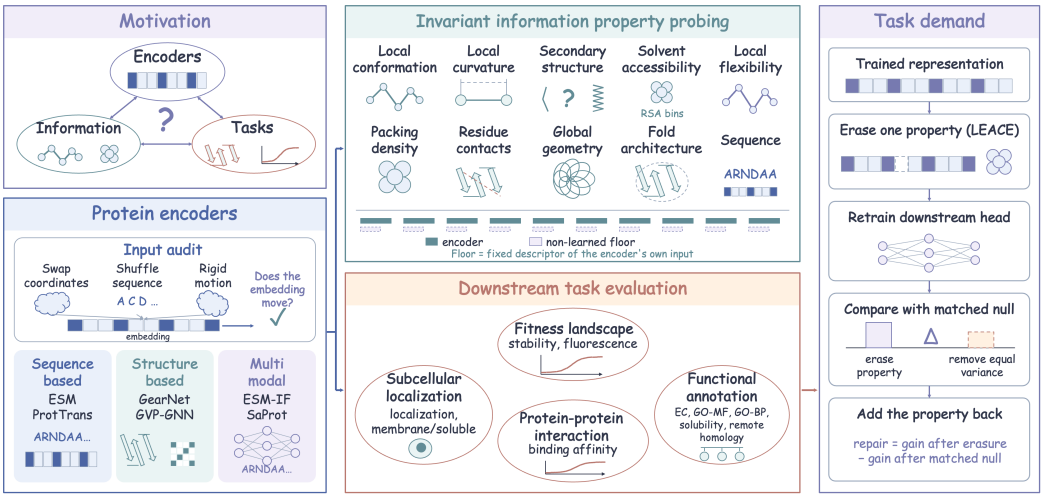


Figure 1: Overview of ProtCompass. An input audit first determines whether each encoder uses sequence, structure, or both. Structural probing compares encoder representations with non-learned descriptors of the same inputs, while downstream evaluation measures performance across function, localization, fitness, and interaction tasks against simple sequence baselines. Task demand is then tested by erasing one property, comparing the resulting loss with a variance-matched null, and adding the property back to determine whether it selectively restores performance.

et al., 2021; van Kempen et al., 2023; Zhang et al., 2022; Hsu et al., 2022; Batatia et al., 2022; Dauparas et al., 2022; Morehead & Cheng, 2024; Birnbaum & Keating, 2026; Fan et al., 2023; Su et al., 2023; Hayes et al., 2025; Kroll et al., 2026). The complete panel is listed in Appendix Table 3.

We classify encoders by the inputs they actually use. For each encoder, we perturb the same 32 proteins by swapping or jittering coordinates, shuffling the sequence, and rotating or translating the structure. Coordinate swapping tests whether the encoder uses structural coordinates, while sequence shuffling tests whether it uses residue identity. Rigid-body motion serves as a sanity check: the representation should remain unchanged when the same protein is only rotated or translated. This audit identifies 10 encoders that do not use coordinates and 23 that do. We exclude EGNN, GeoBPE and dMaSIF because they won't respond to changes of coordinates, rigid-body motion or return null embeddings. For the remaining coordinate-reading encoders, a second audit determines which atom types are actually used so that each model can be compared with a baseline built from the same structural information.

2.3 Protein Properties and Downstream Tasks

Structural probing. The probing suite contains 21 probes grouped into ten protein property groups (PPG, I1–I10), spanning local backbone geometry, secondary structure, solvent accessibility, flexibility, burial and packing, pairwise distances, global shape, fold architecture, and sequence recovery. Fifteen probes enter the matched-floor comparison; the remaining six are alternative formulations or uninformative variants. Definitions are given in Appendix H and Tables 40 and 40.

Downstream prediction. We assembled 14 biological tasks covering structure, function, localization, solubility, fitness, and binding (AlQuraishi, 2019; Rao et al., 2019; Notin et al., 2023; Jamasb et al., 2024; Dallago et al., 2021; Dürr et al., 2023; Jankauskaitė et al., 2018; Thumuluri et al., 2022). Four were excluded after their labels were found unsuitable for the intended evaluation, leaving ten scored tasks. Stability, fluorescence, solubility, and PPI affinity use ESMFold-predicted structures where experimental structures are unavailable. Full definitions, metrics, and splits are given in Appendix H.5.

2.4 Measuring What a Task Uses

The erasure analysis asks whether a downstream task uses a protein property after controlling for the representation capacity removed with that property.

Property erasure. We compute eight protein-level properties for every protein. Six independent properties are erased, giving 60 task–property pairs across ten tasks; charge and hydrophobic fraction are treated as functions of amino-acid composition. For representation X and property p, LEACE (Belrose et al., 2023) removes the linear subspace carrying p. The downstream head is then retrained on the erased representation. The main analysis therefore measures *linear-head demand*: information about p that is linearly accessible in the representation and used by a retrained linear head.

Controls separate property loss from representation loss. Erasing a property can reduce performance for two reasons: the task may use that property, or the erasure may simply remove useful representation capacity. We use two controls to distinguish them. The *sham* control applies the same LEACE procedure to a row-permuted version of the property, preserving the erasure procedure while removing any meaningful relation to the protein. However, the sham typically removes less of the representation than the real erasure. We therefore add a *variance-matched* control that removes the same amount of representation variance as the real erasure, but from directions carrying little of the target property. If the real erasure causes a larger loss than this control, the difference is more specifically associated with removing the property rather than simply removing representation capacity. The variance match is within 1% on 62% of audited cells and within 5% on 84%.

For a retrained-head test score $S(\cdot)$, the erasure effect relative to a set of control erasures $\mathcal{N}$ is

$$\Delta(\mathcal{N}) = \operatorname*{mean}_{E \in \mathcal{N}} S(EX) - S(E_p X).$$

Repair test. To determine whether the loss is specific to p, we add an independently computed descriptor q of the property back to both the erased and null representations and retrain:

$$R(\mathcal{N}) = \big[S(E_p X \oplus q) - S(E_p X)\big] - \operatorname*{mean}_{E \in \mathcal{N}} \big[S(EX \oplus q) - S(EX)\big]. \tag{1}$$

A positive repair contrast means that q restores more performance after its own erasure than after a matched control erasure. We denote the resulting quantities by Δ_{sham}, Δ_{var}, R_{sham}, and R_{var}. A row-permuted descriptor provides the negative control. Sensitivity is calibrated from the loss produced by erasing the task label itself and from the spread of the null distributions (Appendix D.5).

3 Main Results

3.1 Structural Probing: Separating Learned Signal from Input

Structural probes are commonly interpreted as evidence that an encoder has learned protein structure, where 11 of the 15 labels in our comparison are determined directly by the coordinates given to structure encoders (Appendix H.4). We therefore compare each encoder with a fixed, non-learned floor built from the same input it is measured to read and evaluated through the identical probing protocol.

Probe scores on determined labels largely reflect information already present in the input. Each entry in Figure 2 compares one encoder on one probe label with a baseline built from the same input. Across the 245 coordinate-encoder–label comparisons on determined labels, 231 score below their input-matched baseline, 3 are unresolved, and 11 score above it; none of the 110 sequence-only encoder–label comparisons falls below its sequence baseline (Table 1; Appendix B.2). Removing contact density and contact order, whose baselines directly contain the target, leaves the same separation: 186 of 200 coordinate comparisons fall below the baseline, compared with 0 of 90 sequence-only comparisons. A label-agnostic neighbourhood baseline gives 131 of 176 versus 0 of 80 (Appendix B.2). On the five labels whose baseline reaches 0.958 or more, the count follows from the construction; on the other four determined labels, 74 of 88 coordinate comparisons fall below their baselines, including 35 of 48 pretrained comparisons (Appendix B.3). With the same weights, removing the structure input moves SaProt on all eleven determined labels and ESM3 on eight from below the floor of their atoms to above the sequence floors (Appendix C.2).

Undetermined labels show the complementary pattern. If determinacy explains the floor comparison, encoders should exceed the floors when the input does not fix the label. This is what we observe: encoders beat every written floor on relative solvent accessibility, B-factor, and fold architecture (Appendix B.8). These results identify where probing measures predictive information beyond what the input itself provides.

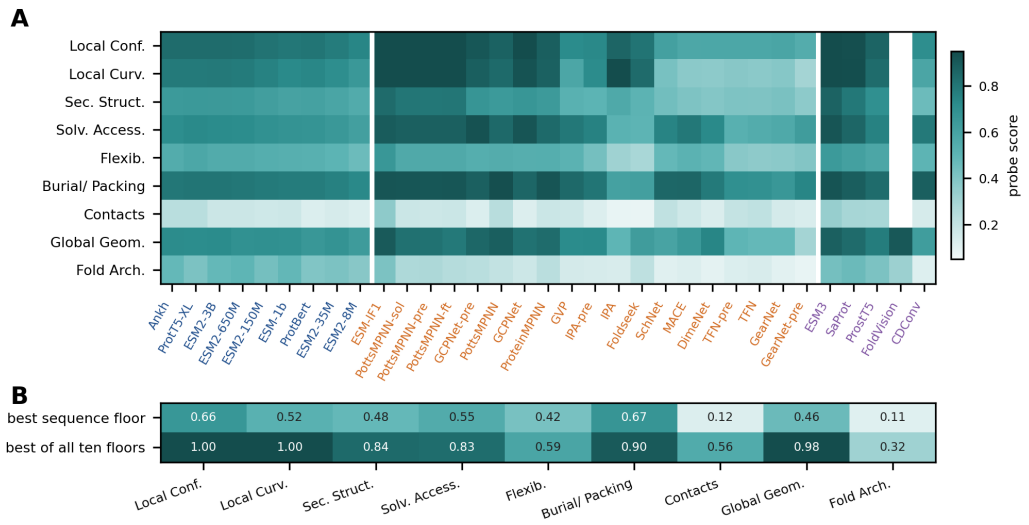


Figure 2: Structural probe scores depend on what the encoder receives as input. (A) Test scores of 33 encoders across nine protein property groups (PPGs), using the readout selected on validation. **(B)** Scores of the corresponding non-learned input floors. Sequence-only encoders remain above their sequence floors, whereas coordinate-reading encoders generally fall at or below floors built from their own atoms on labels determined by those atoms. On labels not determined by the stored input, some encoders exceed every floor. Main-text comparisons use floors matched to the inputs measured for each encoder.

Table 1: Structural probing against input-matched floors. Cells below the floor / evaluated cells, with below defined by a paired 95% confidence interval entirely below zero. Each cell is an encoder and label pair. The first three rows use the superfamily split. Label sets differ across rows; all rows concern coordinate-determined labels. Full results are in Appendices B.2, B.3 and E.1.

Comparison	Labels	Coordinate-reading	Sequence-only
Input-matched floors	11	231 / 245	0 / 110
Excluding contact density and contact order	9	186 / 200	0 / 90
Label-agnostic neighbourhood floor ($K = 8$)	8	131 / 176	0 / 80
Unseen CATH topologies	11	234 / 245	0 / 110

What a probe reveals does not necessarily determine which encoder is most useful in practice. We next ask which encoders perform best on biological prediction tasks.

3.2 Downstream Prediction: What Do Protein Encoders Add?

We evaluate the 33 encoders on ten biological prediction tasks spanning enzyme and Gene Ontology annotation, fold recognition, localization, solubility, and mutation effects on fluorescence, stability and binding affinity (Figure 3; Appendix Table 44). For each encoder–task pair, the readout is selected on validation and its test performance is averaged over three seeds to reduce training noise (median seed standard deviation 0.0059). Training is capped at 6,000 proteins on eight tasks (Appendix H.6).

Sequence-only encoders lead across the full panel. Sequence-only encoders achieve higher mean performance on all ten tasks (Figure 3). The advantage ranges from 0.090 to 0.473 on nine tasks, with a mean gap of 0.186 across the panel; eight of these task-level differences remain significant after Holm correction. PPI affinity is effectively unresolved: the gap is only 0.011, compared with a mean seed standard deviation of 0.101.

The sequence-only advantage narrows among large pretrained encoders. The full-panel comparison is partly confounded by model scale and training, because many coordinate-reading encoders are small or randomly initialized. Restricting both groups to pretrained models with at least 100M parameters leaves eight sequence-only and four coordinate-reading encoders. Sequence-only models score higher on eight of ten tasks, but the average gap shrinks to 0.048 and is not resolved across architecture families ($+0.049$ $[-0.016, +0.114]$; permutation $p = 0.093$; Appendix C.3). The result

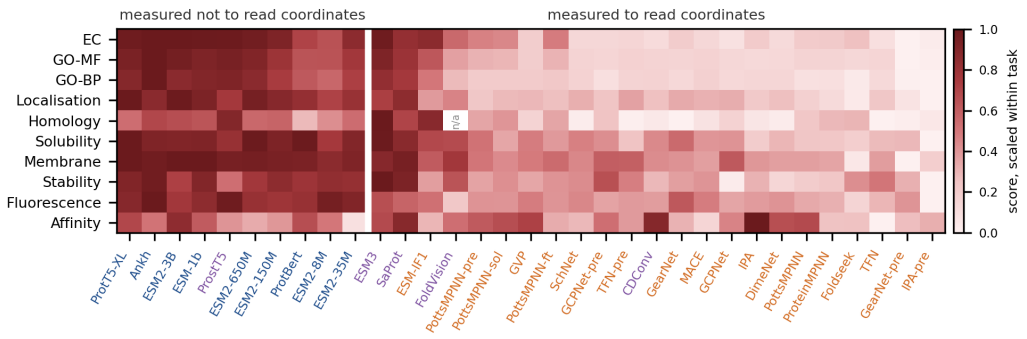


Figure 3: Downstream performance of 33 protein encoders across ten biological prediction tasks. Each cell shows test performance averaged over three seeds for the readout selected on validation; `n/a` marks the one unevaluated encoder–task pair. Color is scaled within each task because the tasks use different metrics (Appendix H.5). Encoders measured not to read coordinates are shown on the left and coordinate-reading encoders on the right, with each group ordered by mean rank across tasks. Name colors indicate declared encoder class: sequence-based (navy), structure-based (orange), and multimodal (purple).

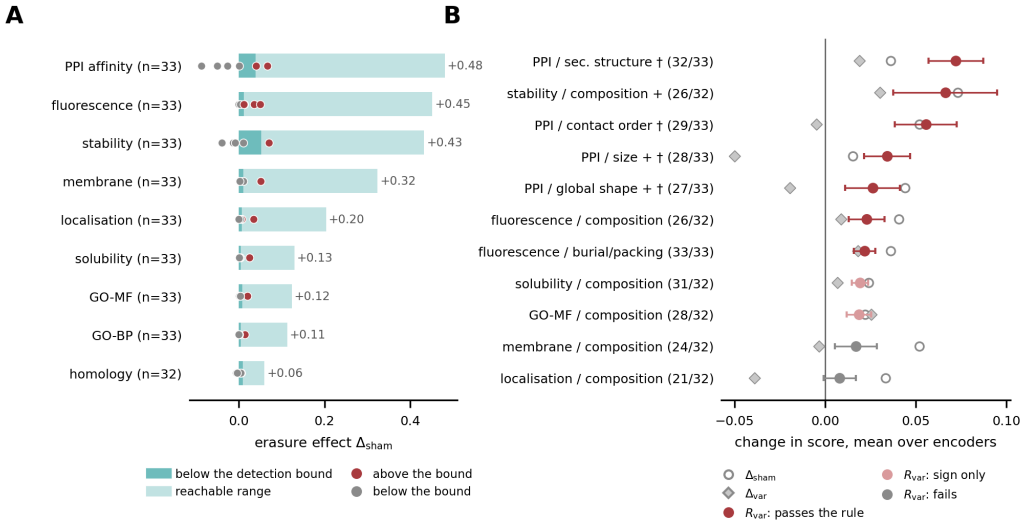


Figure 4: Property use measured by erasure and repair. (A) Loss after erasing each property relative to a sham erasure, with each task's reachable range and minimum detectable effect. **(B)** Candidate task–property pairs compared against a variance-matched null. Open circles show the sham-referenced erasure effect Δ_{sham}, diamonds the variance-matched effect Δ_{var}, and filled points the repair contrast R_{var} with 95% intervals over encoders. Dark points pass the repair rule; † marks PPI pairs that fail with chain-disjoint folds.

is similar after projecting representations to a common dimension (Appendix C.3). Within ESM3 and SaProt, adding the structure input to the same weights raises fold recognition in both and function annotation in SaProt, and lowers fluorescence, solubility and localization in both (Appendix C.2).

Simple sequence features explain much of four tasks. A baseline using only amino-acid composition and length comes within 0.02 of the median encoder on solubility, GO-BP, localization, and fluorescence (Appendix C.1). Large pretrained sequence models still add clear predictive value, exceeding this baseline by 0.10–0.20 on average. Thus, part of the downstream signal is already captured by simple sequence statistics, while stronger representations contribute additional information. We next ask whether downstream models actually use these protein properties.

3.3 Concept Erasure: Separating Property Use from Representation Loss

We test which protein properties downstream models use by removing each property from the representation with LEACE and retraining the prediction head. The erasure removes linear access to

the property: its linear readout falls from a median 0.669 to -0.003, while nonlinear readouts retain much of the information (Appendix D.2). Synthetic tests recover the known property when it is used and give effects within ± 0.007 when it is not (Appendix D.1).

Standard erasure can mix property loss with representation loss. Using the usual sham control, eight task–property pairs show a consistent performance drop (Figure 4A). However, the real erasure removes a median $3.6\times$ more representation variance than the sham. When we instead compare with a control that removes the same amount of variance, only two of the eight effects remain consistent, and one does not reproduce on held-out proteins (Appendices D.4 and E.4). Thus, many apparent erasure effects arise from how much of the representation is removed.

Adding the property back tests property-specific use. We add each erased property back and ask whether it restores more performance after its own erasure than after the variance-matched control. Across 60 task–property pairs, this repair test identifies seven candidates: four on PPI affinity, two on fluorescence, and amino-acid composition on stability (Figure 4B). The same test correctly identifies the relevant property in synthetic examples where the answer is known (Appendix D.7).

Follow-up controls reduce the seven candidates to one. The four PPI signals disappear when training and test sets share no protein chains, linking them to chain reuse in the original split (Appendix D.13). For fluorescence, simple mutation statistics alone reach Spearman $\rho = 0.49$; after including them, the repairs for composition and burial/packing fall to $+0.004$ and $+0.0003$ (Appendix D.10).

Amino-acid composition on stability is the remaining signal. With the same competing features included, composition restores $+0.100$ performance and is positive for 30 of 32 encoders. The signal also remains when variants are compared within each parent protein and under alternative analysis settings, including a kernel-based intervention ($+0.085$ on 32 of 33 encoders; Appendices D.11 and D.12). To find where this signal comes from, we split each variant's composition into two parts: the composition of its parent protein and the change made by the mutation. The parent part restores $+0.085$ and is positive for 29 of 32 encoders, even when the exact mutation, its position and its local structure are also given; the mutation part restores nothing (Appendix D.10).

In conclusion, these results show that erasure should be interpreted relative to how much representation is removed, and property use is most clearly supported when adding that property back selectively restores performance.

3.4 Visualization and Residue-Level Evaluation

Latent-space organization. The latent-space atlas (Appendix K) shows the same ordering without a trained head: on all ten tasks the encoder whose ten nearest neighbors most often share a protein's label reads sequence only (Table 48). Its maps also compare encoder neighbourhoods and expose degenerate embeddings; neighbourhood purity against chance detects all four failures in the discarded embedding cache (Appendix K).

Rankings across prediction levels. The ten biological tasks use pooled protein embeddings. A coordinate encoder may carry residue-resolved geometry, which a pooled readout cannot show. We therefore compare the pooled rankings with residue-pair contact prediction and residue-level interface prediction (Figure 5; Appendix Table 38). Contact labels are determined by the chain's Cα coordinates. Interface labels depend on a partner chain absent from the encoder's input. The per-encoder results and readout protocols remain in Appendix F.

Encoder choice depends on the prediction level and task. On residue-pair contacts, coordinate-reading and sequence-only encoders perform similarly on average (AUPRC 0.263 vs. 0.247, one-sided Mann–Whitney $p = 0.35$), and an encoder's contact-prediction rank is only weakly related to its rank on the ten pooled tasks ($\rho = 0.22$). Several coordinate-reading encoders move sharply upward on contacts; the pooled ranking transfers well to interface-residue prediction ($\rho = 0.73$).

3.5 Ablation Studies

The probing result is robust across analysis choices. We tested whether the separation between coordinate-reading and sequence-only encoders depends on how the baseline is chosen, representation

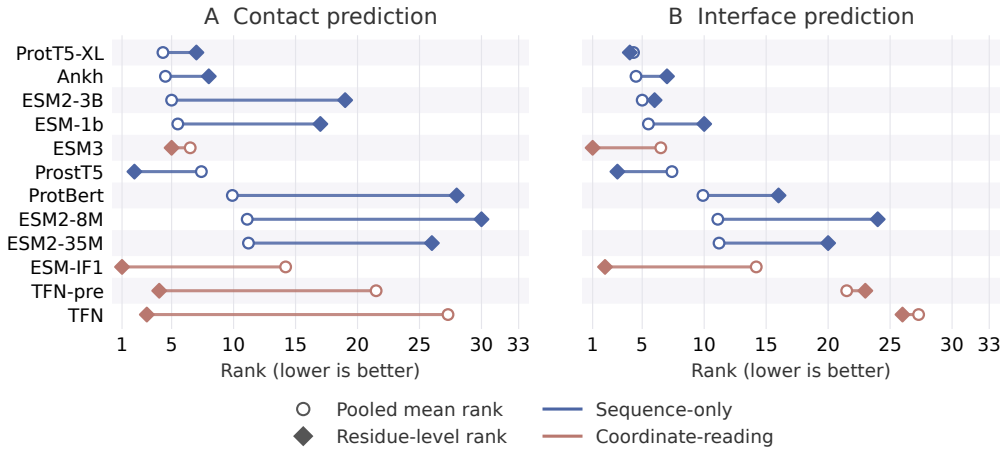


Figure 5: Encoder rankings across prediction levels. Twelve shared entries from the appendix task tables. Circles: mean rank across ten pooled tasks; diamonds: contact or interface rank. Original panel sizes are 33, 32 and 28; lower is better. The contact table selects high ranks and large downward shifts, so this subset illustrates changes without estimating their prevalence. SaProt lacks an interface evaluation.

size, model scale, or the exact input available to each encoder. First, we selected the baseline and best encoder on one half of the test proteins and evaluated them on the other; every comparison kept the same direction (Appendix B.1). We then projected all representations to the same 116-dimensional size, and the separation between the two encoder groups remained (Appendix B.7). We also strengthened the structural and sequence baselines and restricted the comparison to large pretrained models; the main pattern was unchanged (Appendices B.8, B.9, and B.3). Finally, an atom-level analysis showed that the result follows which structural information each encoder actually receives, rather than model size or pretraining alone (Appendix B.3). Within one architecture, pretraining does not move an encoder above its floor on any determined label (Appendix B.4). Together, these checks show that the probing result is robust to the main design choices in the analysis. **Alternative probing controls support the same interpretation.** Standard control-task selectivity gives a different ranking from the input floors on DSSP-3, showing that the two controls answer different questions (Appendix B.5). We also retain the directional information in the vector channels of four equivariant encoders; their probe scores increase, but all 44 determined encoder–label comparisons remain below their input floors (Appendix B.6). Finally, the same separation holds with validation-selected readouts and with a label-agnostic neighbourhood baseline (Appendix B.2).

The downstream comparison is robust to model size and training budget. We tested whether the sequence-only lead is explained by representation width, model scale, training status, or the amount of downstream training data. After projecting all representations to the same 116-dimensional width, the mean sequence-only lead remains 0.164, close to 0.186 at native widths (Appendix C.3). Restricting comparisons by training status, parameter count, and architecture family gives the same overall pattern (Appendix C.3). Reducing the downstream training set to 25% or 50% generally shrinks the lead, indicating that the training cap does not create the observed advantage (Appendix C.3). We also compare all encoders with a composition-and-length baseline to separate gains from simple sequence statistics (Appendix C.1).

The erasure and repair analysis is calibrated across control settings. We first tested whether the intervention correctly identifies property use when the answer is known. Synthetic tasks recover planted properties and rarely assign effects to unrelated ones (Appendices D.1 and D.7). We then verified what LEACE removes, compared our implementation with the reference implementation, and tested whether adding the erased property back selectively restores performance (Appendices D.2, D.3, D.4, and D.8). Sensitivity analyses vary the amount of representation removed, the decision threshold, the encoder subset, and the weighting of architecture families; the overall conclusions remain stable across these choices (Appendices D.5, D.6 and D.11).

Follow-up controls isolate the remaining stability signal. We next tested the candidate task–property relationships against simpler biological explanations. Removing chain overlap between training and test eliminates the four PPI candidates, while accounting for mutation statistics removes both fluorescence candidates (Appendices D.13 and D.10). Amino-acid composition on stability remains under calibrated thresholds and stricter variance matching (Appendix D.11). With a nonlinear head the signal is positive on 7 of 12 encoders, while a kernel-space intervention gives a repair of $+0.085$ on 32 of 33 encoders, showing that its measured strength depends on how the representation is read (Appendix D.12).

3.6 Confirmation on Unseen Data

The determinacy pattern generalizes to unseen folds and new labels. If this interpretation is general, it should hold beyond the data and labels used to develop the floors. On 1,924 domains from unseen CATH topologies, 234 of 245 coordinate cells remain below their floor and no sequence-only cell does; undetermined labels show larger margins by $+0.178$ $[+0.146, +0.217]$ (Appendix E.1). On three new labels whose determinacy and floors were fixed before scoring, every determined family falls below its floor and undetermined families shift upward by $+0.136$ $[+0.081, +0.188]$ (Appendix E.3). Thus, what a probe score measures is systematically shaped by whether the encoder's input already determines the target. Two CATH 4.4 cohorts absent from CATH 4.3 test the probing results on new domains (Appendix E.2). The sequence-only and undetermined-label predictions hold in both; two predictions on individual encoders fail. The second cohort's frozen hypotheses also include input-matched floors and new topologies. These failures limit the encoder-specific claims. Downstream confirmation uses previously unscored proteins, with the fitting and scoring rules frozen before evaluation (Appendix E.4). Selective repair holds on these proteins, while the variance-matched erasure effect can lose its consistency in sign. Paired protein bootstraps quantify uncertainty, and multi-output readouts complete the four tasks requiring larger heads (Appendix E.4). Seed uncertainty for the probing and downstream panels is reported in Appendix J.

4 Related Work

Protein encoders model sequence, structure or both (Lin et al., 2023; Elnaggar et al., 2022; Zhang et al., 2022; Jing et al., 2020; Su et al., 2023; Hayes et al., 2025), and downstream benchmarks assess protein function, structure and mutation effects (Rao et al., 2019; Xu et al., 2022; Notin et al., 2023; Townshend et al., 2020). Probing studies examine recoverable properties (Vig et al., 2020; Prabakaran & Bromberg, 2026), while control tasks and conditional probing assess the information a representation adds beyond a baseline (Hewitt & Liang, 2019; Hewitt et al., 2021). Erasure methods test property use by removing information from a representation (Elazar et al., 2021; Belrose et al., 2023). A comprehensive review of related work is provided in Appendix G.

5 Conclusion

ProtCompass shows that common protein-encoder evaluations can mix learned signal with information supplied by the input or evaluation itself. In structural probing, coordinate encoders often preserve properties already determined by their inputs. In downstream prediction, simple sequence features explain substantial signal, while our large pretrained models show no consistent gain from adding coordinates for pooled tasks. In concept erasure, controlling how much representation is removed eliminates most apparent property effects; after follow-up controls, amino-acid composition on stability remains the single linear-head signal among the tested pairs. These results suggest a simple principle for evaluating protein representations: separate learned signal from input and evaluation effects before assigning biological meaning to a score. Input-based probing baselines, simple sequence controls, and capacity-matched erasure controls make explicit what a representation actually adds.

Limitations. Our conclusions are limited to the encoders, tasks, and readouts evaluated here. The main intervention analysis measures properties accessible to linear heads, several downstream tasks use predicted structures. We have only included important downstream tasks in biological applications.

REFERENCES

Etowah Adams, Liam Bai, Minji Lee, Yiyang Yu, and Mohammed AlQuraishi. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models, February 2025. URL http://dx.doi.org/10.1101/2025.02.06.636901.

Mohammed AlQuraishi. Proteinnet: a standardized data set for machine learning of protein structure, 2019. URL https://arxiv.org/abs/1902.00249.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields, 2022.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances, 2021.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, 2023.

Foster Birnbaum and Amy E. Keating. Beyond native sequence recovery: Improved modeling of the sequence-energy landscape of protein structures. *bioRxiv : the preprint server for biology*, 2026. doi: 10.64898/2026.01.14.699067.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. ProteinBERT: A universal deep-learning model of protein sequence and function. *Bioinformatics (Oxford, England)*, 38(8):2102–2110, February 2022. ISSN 1367-4811. doi: 10.1093/bioinformatics/btac020.

Christian Dallago, Jody Mou, Kadina E. Johnston, Bruce J. Wittmann, Nicholas Bhattacharya, Samuel Goldman, Ali Madani, and Kevin K. Yang. FLIP: Benchmark tasks in fitness landscape inference for proteins. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*. openRxiv, November 2021. doi: 10.1101/2021.11.09.467890.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, October 2022. ISSN 1095-9203. doi: 10.1126/science.add2187.

Nicki Skafte Detlefsen, Søren Hauberg, and Wouter Boomsma. Learning meaningful representations of protein sequences. *Nature Communications*, 13(1), April 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29443-w.

Simon L. Dürr, Andrea Levy, and Ursula Rothlisberger. Metal3D: a general deep learning framework for accurate metal ion location prediction in proteins. *Nature Communications*, 14(1):2713, May 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-37870-6. URL https://www.nature.com/articles/s41467-023-37870-6.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. In *Transactions of the Association for Computational Linguistics*, 2021.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, October 2022. ISSN 1939-3539. doi: 10.1109/tpami.2021.3095381.

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, January 2023.

Hehe Fan, Zhangyang Wang, Yi Yang, and Mohan Kankanhalli. Continuous-discrete convolution for geometry-sequence modeling in proteins. In *The Eleventh International Conference on Learning Representations*, 2023.

Daria Frolova, Marina Pak, Anna Litvin, Ilya Sharov, Dmitry Ivankov, and Ivan Oseledets. Mulan: multimodal protein language model for sequence and structure encoding. *Bioinformatics Advances*, 5(1), December 2024. ISSN 2635-0041. doi: 10.1093/bioadv/vbaf117. URL http://dx.doi.org/10.1093/bioadv/vbaf117.

P. Gainza, F. Sverrisson, F. Monti, E. Rodolà, D. Boscaini, M. M. Bronstein, and B. E. Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, February 2020. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0666-6. URL https://www.nature.com/articles/s41592-019-0666-6.

Zhangyang Gao, Cheng Tan, Yijie Zhang, Xingran Chen, Lirong Wu, and Stan Z. Li. ProteinInvBench: Benchmarking protein inverse folding on diverse tasks, models, and metrics. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. doi: 10.52202/075280-2983.

Zhangyang Gao, Hao Wang, Cheng Tan, Chenrui Xu, Mengdi Liu, Bozhen Hu, Linlin Chao, Xiaoming Zhang, and Stan Z. Li. PFMBench: Protein foundation model benchmark, 2025.

Edith Natalia Villegas Garcia and Alessio Ansuini. Interpreting and steering protein language models through sparse autoencoders, 2025. URL https://arxiv.org/abs/2502.09135.

Johannes Gasteiger, Shankari Giri, Johannes T. Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules, 2020.

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, August 2025. doi: 10.1073/pnas.2506316122.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 2025. doi: 10.1126/science.ads0018.

Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4), September 2024. ISSN 2631-9268. doi: 10.1093/nargab/lqae150.

Pedro Hermosilla and Timo Ropinski. Contrastive representation learning for 3D protein structures, 2022.

Pedro Hermosilla, Marco Schäfer, Matěj Lang, Gloria Fackelmann, Pere Pau Vázquez, Barbora Kozlíková, Michael Krone, Tobias Ritschel, and Timo Ropinski. Intrinsic-extrinsic convolution and pooling for learning on 3D protein structures, 2020.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, 2019.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D. Manning. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. *ICML*, 2022. doi: 10.1101/2022.04.10.487779.

Arian R. Jamasb, Alex Morehead, Chaitanya K. Joshi, Zuobai Zhang, Kieran Didi, Simon V. Mathis, Charles Harris, Jian Tang, Jianlin Cheng, Pietro Lio, and Tom L. Blundell. Evaluating representation learning on the protein structure universe, 2024. URL https://arxiv.org/abs/2406.13864.

Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. SKEMPI 2.0: An updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics (Oxford, England)*, 35(3):462–469, July 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/bty635.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael J. L. Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons, 2020.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of ICML*, 2019.

Alexander Kroll, Shantanu Yadav, and Martin J. Lercher. FoldVision: A compute-efficient atom-level 3D protein encoder, January 2026.

Tim Kucera, Carlos Oliver, Dexiong Chen, and Karsten Borgwardt. ProteinShake: Building datasets and benchmarks for deep learning on protein structures. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 58277–58289. Curran Associates, Inc., 2023. doi: 10.52202/075280-2539.

Francesca-Zhoufan Li, Ava P. Amini, Yisong Yue, Kevin K. Yang, and Alex X. Lu. Feature reuse and scaling: Understanding transfer learning with protein language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pp. 27351–27375, Vienna, Austria, July 2024. JMLR.org.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. ISSN 1095-9203. doi: 10.1126/science.ade2574.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv : the preprint server for biology*, 2021. doi: 10.1101/2021.07.09.450648.

Alex Morehead and Jianlin Cheng. Geometry-complete perceptron networks for 3D molecular graphs. *Bioinformatics (Oxford, England)*, 40(2), February 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae087.

Viet Thanh Duy Nguyen and Truong-Son Hy. Advances in protein representation learning: Methods, applications, and future directions, 2025.

Pascal Notin, Aaron W. Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Hansen Spinner, Nathan Rollins, Ada Shaw, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Rose Orenbuch, Yarin Gal, and Debora S. Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*. openRxiv, December 2023. doi: 10.1101/2023.12.07.570727.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of ACL*, 2020.

R. Prabakaran and Yana Bromberg. Quantifying uncertainty in protein representations across models and tasks. *Nature Methods*, 23(4):796–804, April 2026. ISSN 1548-7105. doi: 10.1038/s41592-026-03028-7.

Maryam Rahimimovassagh, Ivan Garibay, and Niloofar Yousefi. Beyond tokens: Probing higher-order epistasis in learned protein representations, 2026. URL https://arxiv.org/abs/2608.24953.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating protein transfer learning with TAPE, 2019.

Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F. Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. In *Proceedings of Machine Learning Research*. openRxiv, February 2021. doi: 10.1101/2021.02.12.430858.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of ACL*, 2020.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 2019. doi: 10.1101/622803.

K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller. Schnet: a continuous-filter convolutional neural network for modeling quantum interactions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 992–1002, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Siddharth Setlur, Djordje Mihajlovic, and Darrick Lee. Interpreting latent protein language model features with geometric annotations, 2026. URL https://arxiv.org/abs/2608.26419.

Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P Waman, Paul Ashford, Harry M Scholes, Camilla S M Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, Mahnaz Abbasian, Sean Le Cornu, Su Datt Lam, Karel Berka, Ivana Hutařová Varekova, Radka Svobodova, Jon Lees, and Christine A Orengo. CATH: Increased structural coverage of functional space. *Nucleic Acids Research*, 49(D1):D266–D273, November 2020. ISSN 1362-4962. doi: 10.1093/nar/gkaa1079.

Elana Simon and James Zou. InterPLM: Discovering interpretable features in protein language models via sparse autoencoders. *Nature Methods*, 22(10):2107–2117, September 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02836-7.

Pavel Strashnov, Andrey Shevtsov, Viacheslav Meshchaninov, Olga Kardymon, and Dmitry Vetrov. Geommotif: A benchmark for arbitrary geometric preservation in protein generation. *bioRxiv*, 2026. doi: 10.64898/2026.04.18.719378. URL https://www.biorxiv.org/content/early/2026/04/22/2026.04.18.719378.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations*. openRxiv, October 2023. doi: 10.1101/2023.10.01.560349.

Jin Su, Yan He, Shiyang You, Shiyu Jiang, Xibin Zhou, Xuting Zhang, Yuxuan Wang, Xining Su, Igor Tolstoy, Xing Chang, Hongyuan Lu, and Fajie Yuan. A trimodal protein language model enables advanced protein searches. *Nature Biotechnology*, 44(8):1310–1316, October 2025. ISSN 1546-1696. doi: 10.1038/s41587-025-02836-0. URL http://dx.doi.org/10.1038/s41587-025-02836-0.

Yang Tan, Wenrui Gou, Bozitao Zhong, Liang Hong, Huiqun Yu, and Bingxin Zhou. Venusx: Unlocking fine-grained functional understanding of proteins, 2025. URL https://arxiv.org/abs/2505.11812.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds, 2018.

Vineet Thumuluri, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Henrik Nielsen, and Ole Winther. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Research*, 50(W1):W228–W234, July 2022. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkac278. URL https://academic.oup.com/nar/article/50/W1/W228/6576357.

Raphael J. L. Townshend, Martin Vögele, Patricia Suriana, Alexander Derry, Alexander Powers, Yianni Laloudakis, Sidhika Balachandar, Bowen Jing, Brandon Anderson, Stephan Eismann, Risi Kondor, Russ B. Altman, and Ron O. Dror. ATOM3D: Tasks on molecules in three dimensions, 2020.

Darin Tsui, Kunal Talreja, Daniel Saeedi, and Amirali Aghazadeh. Protein circuit tracing via cross-layer transcoders, 2026. URL https://arxiv.org/abs/2602.12026.

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 42(2):243–246, May 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0.

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

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of EMNLP*, 2020.

Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLOS Computational Biology*, 13(1):e1005324, January 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005324. URL http://dx.doi.org/10.1371/journal.pcbi.1005324.

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Dplm-2: A multimodal diffusion protein language model, 2024. URL https://arxiv.org/abs/2410.13782.

Lirong Wu, Yufei Huang, Haitao Lin, and Stan Z. Li. A survey on protein representation learning: Retrospect and prospect, 2023.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Chang Ma, Runcheng Liu, and Jian Tang. PEER: A comprehensive and multi-task benchmark for protein sequence understanding, 2022.

Fei Ye, Zaixiang Zheng, Dongyu Xue, Yuning Shen, Lihao Wang, Yiming Ma, Yan Wang, Xinyou Wang, Xiangxin Zhou, and Quanquan Gu. ProteinBench: A holistic evaluation of protein foundation models, 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), October 2024. ISSN 1091-6490. doi: 10.1073/pnas.2406285121.

Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining, 2022.

Zuobai Zhang, Chuanrui Wang, Minghao Xu, Vijil Chenthamarakshan, Aurélie Lozano, Payel Das, and Jian Tang. A systematic study of joint representation learning on protein sequences and structures, 2023.

REPRODUCIBILITY STATEMENT

Code, embeddings, datasets, weights and every table behind these numbers are released in an anonymous repository (https://anonymous.4open.science/r/ProEmbedEva-C4F6) and a Hugging Face collection (https://huggingface.co/collections/Anonymoususer2223/). Section 2 gives the datasets, splits, probe protocol and input audit; Appendices H.2, H.3, H.5, B and J the probing pipeline, the floors, the tasks, the controls and the seed spread. A latent-space atlas of the proteins of every scored task under every encoder scored on it, with encoders compared by debiased linear CKA (Kornblith et al., 2019), is released with the benchmark (Appendix K).

AI USE STATEMENT

In this work, we used generative AI tools, including large language model assistants, to polish manuscript text, organize material in the Appendix, and assist with debugging code used in the analyses. We also used generative AI tools to help arrange Github repository and huggingface repositories. All AI-assisted outputs were reviewed by the authors, who take full responsibility for the final content of the paper.

Appendix Contents

A The Encoder Panel **18**

A.1 The evaluated panel . . . 18

B Controls for the probing comparison **19**

B.1 Floors matched to each encoder's input, and floors chosen on validation . . . 19

B.2 The floor comparison with one protocol for encoders and floors . . . 21

B.3 The headline counts by label and by encoder . . . 22

B.4 What pretraining changes at a fixed architecture and input . . . 23

B.5 Control-task selectivity against the floors . . . 24

B.6 Equivariant encoders with the direction of their vector channels kept . . . 26

B.7 Is the floor comparison a comparison of capacity? . . . 26

B.8 Labels the input does not determine, and a floor built for fold architecture . . . 26

B.9 The sequence side, built out to match the coordinate side . . . 28

C Controls for the downstream comparison **28**

C.1 A non-learned reference for the downstream tasks . . . 28

C.2 Removing the structure input from the same checkpoint . . . 29

C.3 The group comparison with each confound held fixed . . . 29

D Controls for the erasure analysis **32**

D.1 A positive control: can the protocol find a demand that was planted? . . . 32

D.2 What the erasure removes: linear access . . . 33

D.3 The eraser against the reference LEACE implementation . . . 34

D.4 The null the erasure is read against . . . 34

D.5 What the effect rule can detect, and how its verdicts depend on the settings . . . 36

D.6 The repair rule and the pairs it names . . . 38

D.7 Calibrating the repair rule on labels whose answer is set . . . 40

D.8 Does the named property repair its own erasure? A selective add-back . . . 41

D.9 A fixed head: does the fitted predictor rely on the erased directions? . . . 42

D.10 Competing explanations on fluorescence and stability . . . 43

D.11 The repair rule with task-calibrated floors, well-matched cells and one vote per family . . . 44

D.12 Composition on stability with a nonlinear head and a nonlinear erasure . . . 45

D.13 PPI affinity with chain-disjoint folds . . . 46

E Confirmation on data and splits the development did not use **47**

E.1 A frozen test on folds absent from training . . . 47

E.2 The probing headline on CATH 4.4 domains absent from CATH 4.3 . . . 48

E.3 A pre-registered test of determinacy on new labels . . . 49

E.4 The demand results on downstream proteins never scored . . . 50

F Two residue-level tasks, and what the pooled readout hides **53**

G Extended related work **55**

H Tasks and protocols **56**

H.1 The probing suite 56

H.2 Probe protocol 57

H.3 The non-learned floors, in full 58

H.4 Determined, undetermined, and not established, label by label 58

H.5 The downstream tasks 61

H.6 Downstream results in detail 62

I Tasks and analyses the paper evaluated and does not use **64**

J Seed variability **65**

K Latent-Space Atlas **67**

SCOPE OF THIS APPENDIX

The appendix gives the methods, controls and per-cell results behind the three questions of the main text: what a structural probe measures (Section 3.1), what downstream scores show (Section 3.2) and what a task uses (Section 3.3). Appendix A describes the encoder panel. Appendices B, C and D hold the controls of the probing, downstream and erasure comparisons, Appendix E the confirmation on data the development did not use, and Appendix F two residue-level tasks. Appendix G extends the related work, Appendix H defines the tasks and protocols, Appendix I records the tasks and analyses the paper evaluated and does not use, Appendix J gives the seed spread, and Appendix K the latent-space atlas.

Table 2 lists what each comparison in the paper establishes and what it leaves open.

comparison	reference or intervention	what it establishes	what it does not establish
probe score against floor	fixed descriptor of the audited input, same probe	whether the embedding makes a label easier to read than a written descriptor of the same input (Appendices B.1 and B.2)	that training created the information, or that the input lacks it
random against pretrained	same architecture, input atoms and floor	what pretraining changes in label accessibility (Appendix B.4)	the effect of the input modality
full against sequence-only input	same checkpoint, structure input removed	what the supplied structure changes for probes and tasks (Appendix C.2)	a general effect of structure across architectures
retrained erasure	LEACE against a variance-matched null, head retrained	how much task performance stays reachable after linear access to the property is removed (Appendix D.4)	use by the head fitted on the intact representation
fixed-head erasure	same projections, head fitted on the intact representation	whether the fitted predictor relies more on the erased directions than on matched ones (Appendix D.9)	a unique biological attribution
repair R_{var}	property added back after erasure and after the null	property-specific rescue within the linear head class (Appendices D.6 and D.8)	dependence through nonlinear readouts
stability decomposition	parent composition and substitution vector, each with the other in every arm	which part of the composition signal carries the repair (Appendix D.10)	a physical mechanism of stability

Table 2: What each comparison establishes.

A THE ENCODER PANEL

This appendix lists the encoders the release scores and states how they are grouped, since every comparison in Sections 3.1 and 3.2 is made between groups. The release scores 33 encoders (Table 3). The analyses group them by an input audit that measures how each embedding responds to a coordinate jitter, a coordinate swap, a sequence shuffle and a rigid motion (Section 2), and keep the declared class only as a label. Ten of the 33 are measured not to read coordinates. The declared and measured groups differ for ProstT5, which is declared multimodal and measured sequence-only. Three further encoders were audited and are not released. EGNN declares that it reads coordinates, and its embeddings are measured not to change when the coordinates change. GeoBPE changes its output under a rigid motion by a relative 1.05×10^{-1}; an encoder is called invariant when that change is below $\max(10^{-3}, 2\%$ of the larger of the swap and jitter changes), a threshold that separates numerical noise in reduced precision from frame dependence. dMaSIF returns all-zero embeddings (Appendix K).

A.1 THE EVALUATED PANEL

Table 3 lists the encoders the release scores. Its entries come from the released `meta.json` of every extraction and from `scores2_selected.csv`, as written by `scripts/emit_inventory.py`. The probes of the suite are listed in Table 40 (Appendix H).

Two columns of Table 3 need a note. *Dim.* is the width of the released embedding, read from the stored array, and every number in this paper is computed from that array. It differs from the declared width for two architectures: GVP and GCPNet are listed in their own papers at 148 and 176, and both emit 116 in every released run. The invariant extraction stores each vector channel as one norm, so the width is the number of scalar channels plus the number of vector channels. For GVP this gives $116 = 100 + 16$, where three components per vector channel give the declared 148. The pretrained GCPNet is the official PSR checkpoint, whose 100 scalar channels replace the 128 of the from-scratch configuration, and the same arithmetic gives its 116. *Coord.* is the measured result of the input audit, which places ProstT5 on the sequence-only side. Every encoder is scored on the ten downstream tasks except FoldVision, which has no remote-homology extraction.

Which layer is read. Every embedding is taken from the last representation the encoder computes before any task head or decoder, and where the encoder has residue states a protein is represented by their masked mean over its residues. For the language models this is the final hidden state with special tokens removed; ProstT5 is read on its amino-acid track with the `<AA2fold>` prefix. ESM3 is read at its final trunk embeddings and SaProt at the final hidden states of the residue positions of its combined amino-acid and 3Di tokens. The geometric networks are read at their final invariant node states: GVP and GCPNet keep each vector channel as its norm, and IPA reads the structure in a canonical pose. ProteinMPNN and the PottsMPNN checkpoints are read at the encoder output h_V, before the decoder, and ESM-IF1 at the output of its GVP-Transformer encoder. Foldseek gives a one-hot 3Di state per residue. FoldVision gives only a global embedding and has no residue states.

Table 3: The 33 encoders the release evaluates, grouped by declared class. *Coord.* is what the input audit measures. Parameter counts and widths (*Dim.*) are those of the released checkpoint and the stored array; a dash marks a tokeniser with no learned parameters. Every encoder is scored on the ten downstream tasks except FoldVision, which is scored on nine.

Encoder	Weights	Params.	Dim.	Coord.	Checkpoint
Sequence (9)					
Ankh	pretr.	453M	768	no	`ElnaggarLab/ankh-base`
ESM-1b	pretr.	651M	1280	no	`facebook/esm1b_t33_650M_UR50S`
ESM2-150M	pretr.	148M	640	no	`facebook/esm2_t30_150M_UR50D`
ESM2-35M	pretr.	34M	480	no	`facebook/esm2_t12_35M_UR50D`
ESM2-3B	pretr.	2.8B	2560	no	`facebook/esm2_t36_3B_UR50D`
ESM2-650M	pretr.	651M	1280	no	`facebook/esm2_t33_650M_UR50D`
ESM2-8M	pretr.	8M	320	no	`facebook/esm2_t6_8M_UR50D`
ProtBert	pretr.	419M	1024	no	`Rostlab/prot_bert_bfd`
ProtT5-XL	pretr.	1.2B	1024	no	`Rostlab/prot_t5_xl_half_uniref50-enc`
Structure (19)					
DimeNet	rand.	363K	128	yes	`dimenet`

Table 3 (continued)

Encoder	Weights	Params.	Dim.	Coord.	Checkpoint
ESM-IF1	pretr.	142M	512	yes	`esm_if`
Foldseek	pretr.	–	20	yes	`foldseek_3di`
GCPNet	rand.	1M	116	yes	`gcpnet_official_random`
GCPNet-pre	pretr.	1M	116	yes	`model_epoch_115_localpearson_0_616...`
GVP	rand.	477K	116	yes	`gvp_inproc`
GearNet	rand.	7M	512	yes	`gearnet`
GearNet-pre	pretr.	20M	512	yes	`mc_gearnet_edge.pth`
IPA	rand.	648K	128	yes	`ipa`
IPA-pre	pretr.	648K	128	yes	`best.pt`
MACE	rand.	584K	128	yes	`mace`
PottsMPNN	rand.	2M	128	yes	`pottsmpnn`
PottsMPNN-ft	pretr.	2M	128	yes	`pottsmpnn_ft_pretrained`
PottsMPNN-pre	pretr.	2M	128	yes	`pottsmpnn_pretrained`
PottsMPNN-sol	pretr.	2M	128	yes	`pottsmpnn_soluble_pretrained`
ProteinMPNN	pretr.	2M	128	yes	`proteinmpnn_v48_020.pt`
SchNet	rand.	244K	128	yes	`schnet`
TFN	rand.	251K	128	yes	`tfn`
TFN-pre	pretr.	251K	128	yes	`best.pt`
Multimodal (5)					
CDConv	rand.	619K	256	yes	`cdconv`
ESM3	pretr.	1.4B	1536	yes	`esm3_sm_open_v1.pth`
FoldVision	pretr.	122M	1024	yes	`AlexanderKroll/foldvision-encoder`
ProstT5	pretr.	1.2B	1024	no	`Rostlab/ProstT5`
SaProt	pretr.	651M	1280	yes	`westlake-repl/SaProt_650M_AF2`

B CONTROLS FOR THE PROBING COMPARISON

This section holds the controls behind the probing comparison of Section 3.1: which floors each encoder is held to, the protocol that adds an interval to every gap, the counts by label and by encoder, and the checks that the separation does not come from training status, readout width, pooling, the strength of either side's floors or the choice of encoders. The controls of the downstream comparison are in Appendix C and those of the erasure analysis in Appendix D.

B.1 FLOORS MATCHED TO EACH ENCODER'S INPUT, AND FLOORS CHOSEN ON VALIDATION

Which floors should each encoder be held to? The released rule holds an encoder measured to read coordinates to the best of all ten floors and a sequence-only encoder to the best of the six sequence floors, with the floor and the readout family chosen on the test split. This can overstate a floor in two ways: a floor may be built from inputs the encoder does not receive, and the choice on the test split can favour a floor that is best there by chance. This subsection removes both and arrives at the input-matched rule behind the counts of Section 3.1.

What each encoder reads. The input audit of Section 2 records whether an encoder reads coordinates and residue identity. The coordinate floors are built from particular atoms: `ref_coords` and `ref_pairhist` from CA alone, `ref_backbone` and `ref_global` from N, CA, C, O and CB. An atom-level audit (`protcompass/extract/atom_audit.py`) therefore recomputes the pooled embedding of the same 32 proteins with one atom type moved at a time by Gaussian noise of 1 Å, and with every coordinate scaled by 1.05 about the CA centroid, and counts an input as read when the median relative change exceeds the threshold of the input audit, 10^{-3}. The two groups separate as cleanly as in the input audit: every atom counted as unread changes the embedding by at most 5.2×10^{-7} and every atom counted as read by at least 1.9×10^{-3}. Ten of the 23 coordinate encoders read CA alone, three read N, CA and C (GVP, ESM-IF1 and ESM3), five read N, CA, C and O (ProteinMPNN and the four PottsMPNN runs), two read N, CA and CB (SaProt and Foldseek 3Di), and three read all five atoms (both GCPNets and FoldVision) (Table 4). Those three read the structure file itself, which also holds the side-chain atoms; the audit moves only the five stored atoms, so for them the table is a lower bound on what they read. Seven coordinate encoders do not read residue identity. Scale enters no decision: TFN does not respond to a change of scale, DimeNet and MACE respond just under the threshold, and in any protein consecutive CA atoms sit 3.8 Å apart, so the absolute scale is recoverable from any input that holds the chain.

Table 4: Inputs each encoder is measured to read, and the multi-atom floors it is held to under the input-matched rule. **Scale** is whether the embedding responds to scaling the coordinates by 1.05. Both GCPNets and FoldVision read the structure file, which also holds side-chain atoms the audit does not move.

encoder	residue identity	coordinates	atoms read	scale	multi-atom floors
ESM2-35M	yes	no	–	–	–
ESM2-650M	yes	no	–	–	–
Ankh	yes	no	–	–	–
ProtT5-XL	yes	no	–	–	–
ProstT5	yes	no	–	–	–
ProtBert-BFD	yes	no	–	–	–
ESM-1b	yes	no	–	–	–
ESM2-150M	yes	no	–	–	–
ESM2-8M	yes	no	–	–	–
ESM2-3B	yes	no	–	–	–
TFN	yes	yes	CA	no	rebuilt on CA
DimeNet	yes	yes	CA	no	rebuilt on CA
IPA	yes	yes	CA	yes	rebuilt on CA
MACE	yes	yes	CA	no	rebuilt on CA
GearNet	yes	yes	CA	yes	rebuilt on CA
GearNet-pre	yes	yes	CA	yes	rebuilt on CA
SchNet	yes	yes	CA	yes	rebuilt on CA
IPA-pre	yes	yes	CA	yes	rebuilt on CA
TFN-pre	yes	yes	CA	no	rebuilt on CA
CDConv	yes	yes	CA	yes	rebuilt on CA
GVP	yes	yes	N, CA, C	yes	rebuilt on N, CA, C
ESM3	yes	yes	N, CA, C	yes	rebuilt on N, CA, C
ESM-IF1	no	yes	N, CA, C	yes	rebuilt on N, CA, C
PottsMPNN	no	yes	N, CA, C, O	yes	rebuilt on N, CA, C, O
ProteinMPNN	no	yes	N, CA, C, O	yes	rebuilt on N, CA, C, O
PottsMPNN-pre	no	yes	N, CA, C, O	yes	rebuilt on N, CA, C, O
PottsMPNN-sol	no	yes	N, CA, C, O	yes	rebuilt on N, CA, C, O
PottsMPNN-ft	no	yes	N, CA, C, O	yes	rebuilt on N, CA, C, O
FoldVision	yes	yes	N, CA, C, O, CB	yes	released
GCPNet-pre	yes	yes	N, CA, C, O, CB	yes	released
GCPNet	yes	yes	N, CA, C, O, CB	yes	released
SaProt	yes	yes	N, CA, CB	yes	rebuilt on N, CA, CB
Foldseek 3Di	no	yes	N, CA, CB	yes	rebuilt on N, CA, CB

The input-matched rule. An encoder is held to a floor only if it reads every input the floor is built from; Table 5 calls this eligibility *admissible*. Length and position are treated as inputs of every encoder. The sequence floors other than length and position require residue identity, and the CA floors require CA. The two multi-atom floors are rebuilt from the atoms the encoder reads (`scripts/atom_matched_floors.py`): the pairwise distances of the ± 2 window are taken among those atoms only, the chirality sign of each window offset comes from N, CA and C when the encoder reads all three and from CA($j-1$), CA(j), CA($j+1$) otherwise, and the non-local CA block is unchanged. With all five atoms the rebuilt descriptors equal the released ones exactly, which the script asserts before scoring.

The validation rule. Every floor, the ten released ones and the rebuilt ones, is also scored on the validation split with the same probes, training and early stopping (`scripts/floor_val_scores.py`); only the scored split changes. For each probe and each encoder the floor and its readout family are chosen by validation score among the floors the encoder is held to, and the chosen floor's test score is the one the encoder is compared with. The encoder keeps the readout family of the released rule, the better one on test. That choice can only raise the encoder's score, so every count of cells below the floor under this rule is conservative.

What changes. Table 5 gives the counts under the four rules. Restricting each encoder to floors built from inputs it reads moves two cells above the floor, DSSP-3 for PottsMPNN (pretrained) and DSSP-8 for ESM-IF1, and choosing the floor on the validation split moves none. With both changes, 231 of 245 coordinate-encoder cells fall below the floor and 0 of 110 sequence-only cells do. Length and position are never the binding floor, so treating them as inputs of every encoder decides nothing. The encoder's own readout family is the one remaining choice made on the test split here. Any choice lies between the count with the better family everywhere and the count with the worse

family everywhere, between 231 and 237 of 245 for the coordinate encoders and 0 of 110 for the sequence-only encoders. The protocol of Appendix B.2 makes that choice on validation as well and adds an interval to every gap.

Table 5: Determined-probe cells at or below the floor under the four rules (`scripts/floor_rules.py`). **Eligibility** is which floors an encoder is held to: by input group (the released rule) or by the inputs it is measured to read (the input-matched rule, Table 4).

rule	eligibility	floor chosen on	coordinate encoders	sequence-only encoders
released	input group	test	233 of 245	0 of 110
admissible	measured inputs	test	231 of 245	0 of 110
validation	input group	validation	233 of 245	0 of 110
both	measured inputs	validation	231 of 245	0 of 110

Choosing the floor and the best encoder on other proteins. Under the released rule the floor of each probe is the best of ten floors, chosen on the same test set the gap is reported on. To measure what that choice does, we split the test set in half by protein, choose both the floor and the best encoder on one half, and report their scores on the other, with all 41 sources rescored on both halves. The first half names a different floor on 1 of the 15 probes, no probe changes the sign of its gap, the largest change in any gap is 0.044, and the headline count is the same under both choices: no encoder beats the floor on six of the nine informative determined probes. The bias from choosing the floor on test is therefore smaller than the gaps the paper reports.

B.2 The floor comparison with one protocol for encoders and floors

Does the floor count depend on choosing a readout on the test split? This rerun produces the main-text count of Section 3.1, in which 231 of 245 coordinate-encoder cells fall below the input-matched floor and none of 110 sequence-only cells does. The released sweep scores each source on the test split and takes the better of the two readout families there, for encoders and floors alike. The validation-chosen floors of Appendix B.1 make that choice on validation for the floor and keep it on test for the encoder. This rerun makes it on validation for both (`scripts/symmetric_run.py`, `symmetric_decide.py`). Every source, the 33 encoders and every floor, is refitted with the released protocol: the same seed-0 draw of 3,000, 500 and 1,000 domains for the residue and pair probes and every domain of the split for the protein probes, the same ridge penalty grid, and the same MLP-2 with early stopping on validation. Each fitted readout predicts both the validation and the test split, and the rerun reproduces the released test scores to within 10^{-8}. The floor of a cell is the (descriptor, family) with the best validation score among the floors built only from inputs the encoder is measured to read. The gap between the encoder and that floor gets a 95% percentile interval from 2,000 bootstrap draws of test proteins, with the encoder and the floor scored on the same draw. A cell is *below* the floor when the interval lies below zero, *above* when it lies above zero, and *unresolved* otherwise.

Of the 245 coordinate-encoder cells on the eleven determined labels, 231 are below the input-matched floor, 3 are unresolved and 11 are above; none of the 110 sequence-only cells is below, and all 110 are above (Table 6). At 99.9%, 230 coordinate cells are below and 109 sequence-only cells above. Holding each coordinate encoder to all ten floors within this protocol gives 232 below by interval and 233 by point estimate. The released rule, which also holds coordinate encoders to all ten floors and reads every score on test, gives the same 233. The counts therefore do not depend on choosing the encoder's readout on test. The eleven cells above the floor are ESM3 and ESM-IF1 on DSSP-3, DSSP-8 and core/surface, and the five MPNN networks on core/surface (one of them randomly initialised), by $+0.013$ to $+0.069$; the three unresolved cells are three PottsMPNN checkpoints on DSSP-3. The smallest sequence-only margin is ESM-2 8M on Cα distance bins, $+0.003$ $[+0.000, +0.005]$. On the undetermined labels the rerun gives the margins of Section 3.1 with intervals: ESM3 is above every one of the ten written floors by $+0.065$ $[+0.063, +0.067]$ on RSA and $+0.107$ $[+0.104, +0.109]$ on its binned form, and ESM-IF1 by $+0.073$ $[+0.061, +0.088]$ on B-factor.

A floor family generated by one rule. The ten floors were written one after another, each after the encoders had beaten the previous one. The neighbourhood rows of Table 6, which Section 3.1 reports,

Table 6: The floor comparison on the eleven determined labels under one protocol: readout family of encoders and floors, and the floor itself, chosen on validation; below and above by the 95% interval of the gap over 2,000 paired bootstrap draws of test proteins. Input-matched: floors built only from inputs the encoder reads. Released eligibility: a coordinate encoder held to all ten floors. Neighbourhood: the label-agnostic family of Appendix B.2, on the eight residue and pair labels. Point below: the point estimate of the gap is negative (`symmetric/counts_*.csv`).

		coordinate encoders					sequence-only encoders		
split	floors	cells	point below	below	unresolved	above	cells	below	above
superfamily	input-matched	245	231	231	3	11	110	0	110
superfamily	released eligibility	245	233	232	3	10	110	0	110
superfamily	neighbourhood, $K = 4$	176	129	127	2	47	80	0	80
superfamily	neighbourhood, $K = 8$	176	131	131	1	44	80	0	80
superfamily	neighbourhood, $K = 16$	176	131	126	7	43	80	0	80
topology	input-matched	245	234	234	0	11	110	0	108
topology	released eligibility	245	235	235	0	10	110	0	108
topology	neighbourhood, $K = 4$	176	129	129	1	46	80	0	80
topology	neighbourhood, $K = 8$	176	132	132	0	44	80	0	80
topology	neighbourhood, $K = 16$	176	130	128	4	44	80	0	80

use a family generated by one rule that names no label. For the atom set an encoder is measured to read, grouped by element, the descriptor of a residue holds the K nearest atoms of other residues with their distances and signed sequence separations, and the number of atoms of other residues within 3.5, 5, 7.5, 10 and 13 Å, for $K = 4$, 8 and 16 (`ref_nbr`K`@`S; the $K = 8$ member is the generic floor frozen for the pre-registered test of Appendix E.3). An encoder is held only to the family on its own atom set, to length and position, and, if it reads residue identity, to the one-hot residue and its ± 8 window. The family is residue-level, so it covers the eight residue and pair labels among the determined ones. It places 127, 131 and 126 of the 176 coordinate cells on these labels below the floor at $K = 4$, 8 and 16, against 162 for the written floors on the same cells, and none of the 80 sequence-only cells at any K. The cells it misses are concentrated on Cα distance bins (13 of 22 not below at $K = 8$), a pair label that a per-residue neighbourhood does not describe, and on the backbone dihedrals and core/surface (21 of 66). The separation between the two groups therefore holds with floors that were not designed against any label, and the written floors tighten it.

B.3 The headline counts by label and by encoder

How much of the count of Section 3.1 follows from how the floors are built, and is the separation by input a proxy for scale or pretraining? Table 8 splits the cells of the input-matched rule by label and by encoder group (`scripts/headline_reanalysis.py`, which reads the cells of Appendix B.2 and fits nothing). On the five determined labels whose best floor scores 0.958 or more (ϕ, ψ, local curvature, radius of gyration and asphericity), all 112 coordinate cells are below the floor, and on the two labels a floor contains all 45 are. On these seven the count follows from the construction. On the other four (DSSP-3, DSSP-8, core/surface and Cα distance bins), 74 of 88 coordinate cells are below, 35 of the 48 cells of pretrained coordinate encoders, and 6 of the 12 cells of the three pretrained coordinate encoders of at least 100M parameters that give residue embeddings. No sequence-only cell falls below its floor on any label.

A regression of the gap on the encoder, over the 281 cells of the nine determined labels no floor contains, with label fixed effects and standard errors clustered by the 20 architecture families, gives -0.258 $[-0.370, -0.147]$ for reading coordinates, $+0.045$ $[+0.005, +0.084]$ per tenfold increase in parameters and $+0.084$ $[+0.012, +0.156]$ for pretraining. Scale and pretraining raise the gap, and reading coordinates lowers it with both held fixed, so the separation of the two groups is not a proxy for either. The coefficient on reading coordinates also carries the difference in floor strength between the groups, which determinacy predicts: a floor built from the atoms can compute these labels, and a floor built from the sequence cannot.

Averaged over the nine informative determined probes, the gap of each encoder to its own floor runs from $+0.267$ (ProstT5) to -0.498 (GearNet-pretrained). It rises with scale ($\rho = 0.82$ with log parameters over the panel), and within the structure-based and multimodal classes randomly

initialised networks sit further below their floors than pretrained ones (-0.316 against -0.146). The large pretrained coordinate encoders still fall below the floor on these labels. Under the released rule, the four with more than 100M parameters (ESM3, SaProt, ESM-IF1 and FoldVision) are below it in 24 of 29 cells on the nine informative determined probes. The five exceptions are on DSSP-3, DSSP-8 and core/surface, at margins of $+0.018$ to $+0.031$, the labels on which coordinate encoders pass the floor over the whole panel.

The count above assigns determinacy from the five stored atoms. Assigned from the atoms each encoder reads (the atom-level audit of Appendix B.1), with one rule for every label, a cell is *determined* when the label is an exact closed-form function of those atoms, *not established* when the label is recoverable only through ideal geometry (an ideal carbonyl oxygen for DSSP from N, Cα and C, or a virtual Cβ for core/surface), and *not determined* when an atom the label needs is not read. Of the 200 coordinate cells, 126 are determined, and of these 123 are below the floor, 3 are unresolved and none is above; 58 are not determined, all below; 16 are not established, 5 below and 11 above (Table 7). Every cell above the floor is one in which the encoder's atoms fix the label only through ideal geometry: ESM3 and ESM-IF1 on DSSP-3, DSSP-8 and core/surface, and five MPNN networks on core/surface. One of the five is the randomly initialised PottsMPNN, whose margin ($+0.039$ $[+0.038, +0.041]$) is the largest of the eleven; its featurisation builds the virtual Cβ and the neighbour graph that core/surface depends on, so passing the floor there needs no training. In this panel the atom set an encoder reads goes with its architecture, size and pretraining (the ten Cα-only encoders are geometric networks of 20M parameters or fewer, seven of them randomly initialised), so this assignment does not separate determinacy from encoder identity. The not-established category is the one the pre-registration defined before scoring its new labels (Appendix E.3); applying it to the eleven released labels was done after their cells were known.

Table 7: The 200 coordinate cells on the nine determined labels no floor contains, by the determinacy of the label for the atoms each encoder reads, and with a practical-equivalence margin of 0.01; a cell whose interval crosses ± 0.01 without lying inside it is counted in no margin column (`scripts/headline_reanalysis.py`; existing cells, nothing refitted).

status for the atoms the encoder reads	cells	below	unresolved	above
determined	126	123	3	0
not established	16	5	0	11
not determined	58	58	0	0
with a practical-equivalence margin of 0.01 (coordinate encoders)				
labels	cells	below by > 0.01	within ± 0.01	above by > 0.01
best floor ≥ 0.958	112	101	8	0
the other four	88	72	4	11

On the undetermined labels, ESM3 was pretrained with secondary-structure and solvent-accessibility tracks, and the encoders that read side chains receive the atoms relative solvent accessibility is computed from. With both set aside, ESM-IF1 beats every one of the ten written floors on relative solvent accessibility ($+0.049$), its binned form ($+0.071$) and B-factor ($+0.073$), and seven, seven and two encoders beat every written floor on the three labels. ESM-IF1 was itself trained on CATH 4.3 structures, so its margins may include recall of the probing domains; the sequence models that pass these floors are exposed only through their sequence corpora. On fold architecture the largest margin over every written floor is $+0.181$ on the superfamily split (ESM-2 650M) and $+0.078$ on the topology-disjoint split, where ESM-2 650M does not clear the pair-histogram floor (Appendix B.8).

B.4 What pretraining changes at a fixed architecture and input

Does the separation of Section 3.1 follow the input or the training? Five architectures appear in the panel in a randomly initialised and a pretrained version: GCPNet, GearNet, IPA, TFN and PottsMPNN (with the soluble and fine-tuned PottsMPNN checkpoints as secondary pairs that share the same random reference). Within a pair the architecture, the atoms read and the input-matched floor are the same, so the difference between the two versions is what pretraining changes; for GearNet the random network is a smaller configuration (7M against 20M parameters), so size also differs. Each version keeps its own validation-chosen readout, and the difference on each of

Table 8: The cells of the input-matched rule on the eleven determined labels, split by label and encoder group, and the gap regressed on the encoder (label fixed effects, family-clustered 95% intervals).

labels	encoders	cells	below	unresolved	above
best floor ≥ 0.958 (5)	all coordinate	112	112	0	0
best floor ≥ 0.958 (5)	pretrained coordinate	62	62	0	0
best floor ≥ 0.958 (5)	pretrained coordinate, $\geq$100M	17	17	0	0
best floor ≥ 0.958 (5)	sequence-only	50	0	0	50
label in a floor (2)	all coordinate	45	45	0	0
label in a floor (2)	pretrained coordinate	25	25	0	0
label in a floor (2)	pretrained coordinate, $\geq$100M	7	7	0	0
label in a floor (2)	sequence-only	20	0	0	20
the other four	all coordinate	88	74	3	11
the other four	pretrained coordinate	48	35	3	10
the other four	pretrained coordinate, $\geq$100M	12	6	0	6
the other four	sequence-only	40	0	0	40
gap regressed on the encoder, 281 cells, 20 family clusters, label fixed effects					
log10 parameters		$+0.045$ $[+0.005, +0.084]$			
pretrained		$+0.084$ $[+0.012, +0.156]$			
reads coordinates		-0.258 $[-0.370, -0.147]$			

the fifteen probes of the floor comparison carries a paired protein bootstrap interval (2,000 draws, as in Appendix B.2; `scripts/ab_train_pairs.py`; specification fixed before the run in `symmetric/ab_train_pairs/PRESPEC.md`).

Pretraining moves the probe scores in both directions (Table 9, Figure 6). Over the five primary pairs, 38 of the 75 probe cells rise, 29 fall and 8 are unresolved. It does not move an encoder above its floor on a determined label: 54 of the 55 determined cells of the random versions and 53 of the 55 of the pretrained versions are below the floor. The cells that are not below are PottsMPNN's: core/surface in both versions and DSSP-3 in the pretrained one. On core/surface the random network is above the floor (0.937 against 0.897) and scores higher than the pretrained one (0.910; difference -0.027 $[-0.028, -0.025]$), and on DSSP-3 pretraining raises the score by $+0.075$ $[+0.073, +0.078]$ to a level the interval cannot separate from the floor. The largest rises are IPA on relative solvent accessibility ($+0.305$ $[+0.299, +0.312]$) and core/surface ($+0.197$), two labels its Cα-only input does not determine; the largest falls are IPA on local curvature (-0.229) and ψ (-0.200) and GearNet on the two shape labels (-0.18). On the downstream tasks pretraining helps PottsMPNN and TFN on most tasks, lowers GearNet and IPA on most tasks, and changes GCPNet mainly on stability ($+0.549$); these differences carry only seed spreads, because per-protein predictions of the random networks were not stored. The per-probe and per-task differences are released as tables (`symmetric/ab_train_pairs/train_pairs_matrix.tex` and `train_pairs_downstream.tex`).

The intervals resample test proteins and do not include probe refitting, so on residue-level labels they are narrow, and several differences below 0.01 count as rises or falls. The result bounds what pretraining does in this panel: at a fixed architecture and input it changes how accessible each label is, in both directions, and it does not move an encoder above the floor of its input on a label that input determines.

B.5 Control-task selectivity against the floors

Does the usual guard against a probe that reads a lookup table from its input give the same verdict as the floors? Section 3.5 states that it gives a different ranking. The guard is the control task (Hewitt & Liang, 2019): a control task maps each input type to a random label, and selectivity is the probe's score on the real task minus its score on the control. We built the control task for DSSP-3 by mapping each amino-acid type to a DSSP-3 class drawn once from the training class frequencies, and scored it in the same rerun (Table 10).

Table 9: Pretrained minus randomly initialised version of the same architecture on the fifteen probes of the admissible floor comparison (superfamily split). A probe counts as raised or lowered when the protein-level paired 95% bootstrap interval (2,000 draws) of the difference lies entirely above or below zero. Below floor: determined-label cells (of 11) whose 95% interval against the input-matched floor lies below zero, for the random and the pretrained version. Median Δ: over the fifteen probes.

	Probes (of 15)				
Pair (random → pretrained)	raised	lowered	unresolved	below floor	median Δ
GCPNet	10 (6)	4 (4)	1	11 → 11	+0.010
GearNet	5 (2)	10 (9)	0	11 → 11	−0.025
IPA	9 (6)	5 (5)	1	11 → 11	+0.072
TFN	5 (4)	4 (2)	6	11 → 11	−0.001
PottsMPNN	9 (5)	6 (6)	0	10 → 9	+0.026
PottsMPNN-soluble	9 (5)	6 (6)	0	10 → 9	+0.026
PottsMPNN-ft	9 (5)	6 (6)	0	10 → 9	+0.027
GCPNet-vframe	6 (4)	5 (5)	4	11 → 11	+0.003

Numbers in parentheses: of which determined labels (11). GearNet pairs a 7M random configuration with the 20M pretrained model. PottsMPNN-soluble, PottsMPNN-ft and GCPNet-vframe (frame-projected re-extraction) are secondary pairs.

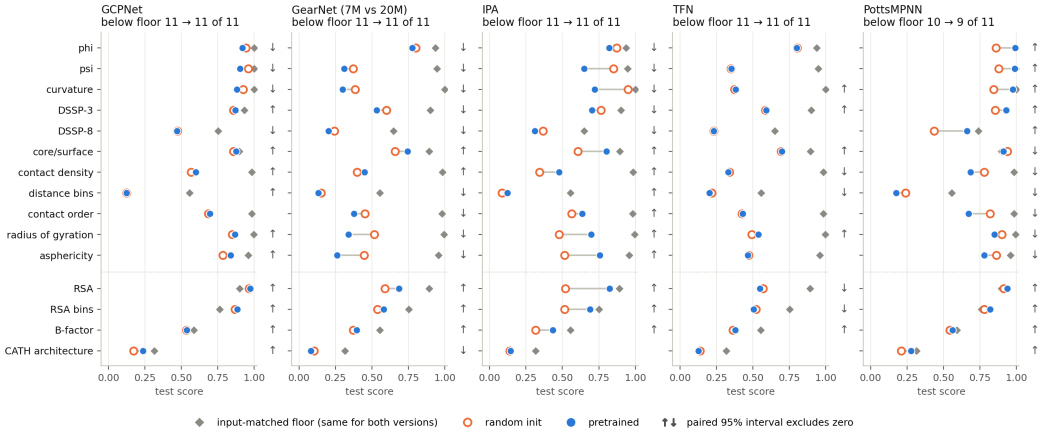


Figure 6: Floor, randomly initialised and pretrained versions of five architectures on the fifteen probes of the floor comparison. Grey diamonds: the input-matched floor, which is the same for both versions; open circles: random initialisation; filled circles: pretrained. Arrows mark a paired 95% bootstrap interval of the difference that excludes zero. Labels below the dotted line are not determined by the stored atoms.

Table 10: Selectivity on DSSP-3 against the floor verdict on DSSP-3 (medians over encoders; floor verdicts from the input-matched rule of Table 6; `symmetric/selectivity_superfamily.csv`).

encoders	input	n	DSSP-3	control task	selectivity	below floor	above floor
coordinate	no residue identity	7	0.931	0.518	+0.412	3	1
coordinate	reads residue identity	15	0.644	1.000	−0.282	14	1
sequence-only	reads residue identity	10	0.828	0.999	−0.172	0	10

Selectivity ranks the encoders in the opposite order to the floors. An encoder that reads residue identity scores close to 1 on the control task, whose label is a function of residue type, so its selectivity is negative: -0.17 at the median for the ten sequence-only encoders, all ten above their floor on DSSP-3, and -0.28 for the fifteen coordinate encoders that read residue identity, 14 of them below their floor. The seven coordinate encoders that do not read residue identity score 0.52 at the median on the control and have the highest selectivity, $+0.41$, and one of them, ESM-IF1, is above its floor on DSSP-3. Selectivity measures whether a probe can use a lookup on residue type. The floors

measure whether a probe score exceeds what the encoder's input already gives, and on this panel the two readings disagree.

B.6 Equivariant encoders with the direction of their vector channels kept

The released extraction stores each equivariant vector channel as its norm, which keeps the representation invariant and discards the channel's direction. We re-extracted the four panel encoders whose final node state holds vector channels, GVP, both GCPNets and MACE, with every vector channel projected on its residue's backbone frame (x along C$\alpha \to$C, z normal to the N, Cα, C plane), which is invariant to rigid motion and keeps the direction of each channel relative to its residue (`extract/extract_vector_frame.py`). TFN and IPA hold no vector channel in their final node state, and ESM-IF1 already rotates its vectors into residue frames. Scored with the protocol of Appendix B.2 against the same input-matched floors, the frame-projected encoders score higher on the determined labels by a median of $+0.009$ (GVP), $+0.013$ and $+0.017$ (GCPNet, pretrained and random) and $+0.087$ (MACE, whose released output uses its scalar channels only), and all 44 of their determined cells stay below the floor, as under the released extraction. On the undetermined labels MACE moves above the floor on RSA and its binned form, and GVP on RSA bins becomes unresolved.

B.7 Is the floor comparison a comparison of capacity?

Does the floor comparison measure width? A fixed descriptor and an embedding are read by the same probe, so the wider and better-conditioned of the two has an advantage unrelated to what it carries, and the floors run from 2 to 357 features and the embeddings from 20 to 2560. A regression over the released runs shows that width raises probe score (median slope $+0.062$ per log unit) and that removing it leaves the sign of the gap unchanged on 11 of the 15 probes; the four that change sit at 0.99 or above for both sides (backbone ϕ, ψ and local curvature) or are B-factor, where the encoders lead in any case (`scripts/capacity_control.py`; released). The rerun below tests the question directly.

Every source at a common width. Every source, the ten floors and the 33 encoders, is projected to 116 dimensions, the width of the narrowest encoders after Foldseek (GVP and GCPNet), before probing, and the whole suite is re-scored under the identical protocol (`PROJECT_K=116 scripts/run_probing_suite.py`; Table 11). The projection is narrower than six of the ten floors and than every sequence model, so it costs the floors and the wide sequence models more than it costs the 116- to 128-wide geometric networks: the wide sequence models lose 0.02 to 0.05 in mean probe score, `ref_pairhist` loses 0.016, and the geometric networks are unchanged or slightly up. The comparison is unchanged. Coordinate encoders fall below the floor built from their own atoms on 237 of 245 cells (233 in the released run), the floor wins on seven of the nine informative determined probes (six in the released run, with DSSP-8 added and DSSP-3 a tie at $+0.001$), and 0 of 110 sequence-only cells fall below the sequence floor, at margins from $+0.004$. On the three undetermined labels and on fold architecture an encoder still beats every floor: RSA $+0.072$, RSA bins $+0.118$, B-factor $+0.049$, and fold architecture $+0.085$, where the projection cuts the released $+0.181$ by more than half and moves the leading encoder from ESM2-650M to ESM3.

B.8 Labels the input does not determine, and a floor built for fold architecture

If the floor comparison follows determinacy, encoders should pass the floors on labels the stored input does not fix. On the three undetermined labels they pass every one of the ten written floors: under the protocol of Appendix B.2, ESM3 by $+0.065$ $[+0.063, +0.067]$ on relative solvent accessibility and $+0.107$ $[+0.104, +0.109]$ on its binned form, and ESM-IF1 by $+0.073$ $[+0.061, +0.088]$ on B-factor. Appendix B.3 gives the margins with ESM3 set aside and the training data behind them.

Fold architecture, whose determinacy is not established, has the largest margin and so most needs a floor built for this label. The standard non-learned way to assign a fold is structural alignment against the training set, and two such floors are added here (`scripts/architecture_alignment_floor.py`), with nothing fitted on the labels. *Alignment 1-NN*: Foldseek in TM-align mode searches each test domain against the training domains of

Table 11: The floor comparison with every source projected to 116 dimensions before probing, next to the released run. Counts are encoders below the floor built from their measured input; for the four labels the stored atoms are not shown to determine, the count is of encoders above the best floor, and the last column is the best encoder minus the best floor at 116.

probe	coord. floor, released → at 116	coord. below, released → at 116	seq. below at 116	best enc. − floor at 116
backbone ϕ	1.000 → 1.000	22 / 22 → 22 / 22	0 / 10	-0.007
backbone ψ	1.000 → 1.000	22 / 22 → 22 / 22	0 / 10	-0.006
local curvature	1.000 → 1.000	22 / 22 → 22 / 22	0 / 10	-0.012
DSSP-3	0.931 → 0.930	18 / 22 → 21 / 22	0 / 10	+0.001
DSSP-8	0.751 → 0.748	21 / 22 → 22 / 22	0 / 10	-0.047
core/surface	0.897 → 0.898	15 / 22 → 15 / 22	0 / 10	+0.039
contact density (floor contains the label)	0.984 → 0.991	22 / 22 → 22 / 22	0 / 10	-0.197
CA distance bins	0.556 → 0.493	22 / 22 → 22 / 22	0 / 10	-0.172
contact order (floor contains the label)	0.982 → 0.982	23 / 23 → 23 / 23	0 / 10	-0.119
radius of gyration	0.996 → 0.997	23 / 23 → 23 / 23	0 / 10	-0.058
asphericity	0.958 → 0.959	23 / 23 → 23 / 23	0 / 10	-0.086
RSA (undetermined)	0.900 → 0.900	10 / 32 above → 9 / 32 above	–	+0.072
RSA bins (undetermined)	0.764 → 0.765	10 / 32 above → 10 / 32 above	–	+0.118
B-factor (undetermined)	0.587 → 0.586	3 / 32 above → 2 / 32 above	–	+0.049
fold architecture (not established)	0.316 → 0.348	14 / 33 above → 10 / 33 above	–	+0.085

the same split, and the architecture of the best hit is the prediction. The structures are written from the coordinates in `labels.h5`, so the floor sees exactly the atoms the encoders were given, and 1,519 of 1,548 test domains get a hit on the superfamily split (median TM-score 0.74) and 1,610 of 1,632 on the topology split (0.69). *Spectral floor*: the sorted eigenvalue spectrum of the CA distance matrix and of the contact map at three cutoffs, length-normalised, read through the same probes, splits, readouts and seeds as every other floor.

Table 12: Fold architecture on the eleven shared CATH architectures, against the descriptor floors and the two floors built for this label. Every number is balanced accuracy, averaged over the seeds of the best readout, the metric and aggregation Section 3.1 uses for this probe.

split	best of the eight floors	alignment 1-NN	spectral floor	best encoder
superfamily-disjoint	0.428 (ref_pairhist)	0.518	0.198	0.628 (esm_if1)
topology-disjoint	0.279 (ref_pairhist)	0.215	0.180	0.343 (prostt5)

On the superfamily-disjoint split, over the eleven shared CATH architectures of Table 12, the alignment floor is the strongest floor we have: at 0.518 it is well past the best descriptor floor, `ref_pairhist` at 0.428, and it takes the best encoder's margin from +0.200 to +0.110. An encoder still passes it. Across the class sets and readouts we ran, the margin over the descriptor floors on this split is +0.18 to +0.32 (+0.181 on the full class set, by ESM-2 650M), and which encoders lead depends on the class set: sequence-only models on the full set, coordinate models on the eleven shared architectures. The spectral floor reaches 0.198 and 0.180 on the two splits, behind both other floors.

On the topology-disjoint split the alignment floor is weak, 0.215 against 0.279 for `ref_pairhist`, because the split removes the structurally similar domains a nearest-neighbour lookup needs. Over 15 seeds the best encoder is ahead of the best floor by +0.078 [+0.057, +0.099] with `ref_pairhist` among the floors and by +0.160 without it; Table 12 averages over seeds before taking the best encoder, which gives the smaller $0.343 - 0.279 = +0.064$. A maximum over 33 encoders is biased upward, so three margins against `ref_pairhist` were fixed in advance: ESM-IF1 +0.055 [+0.034, +0.077], ESM3 +0.056 [+0.032, +0.081] and ProstT5 +0.048 [+0.021, +0.076], while ESM-2 650M, at −0.004 [−0.023, +0.014], does not clear the floor. These pre-specified margins are the numbers to take from this probe. A stronger floor can shrink any of these margins, as the alignment floor does, which is why every floor comparison is read as a lower bound on what the input allows.

B.9 THE SEQUENCE SIDE, BUILT OUT TO MATCH THE COORDINATE SIDE

Does the sequence-only result depend on the sequence floors being weaker than the coordinate floors? The four simplest sequence floors, `ref_length`, `ref_position`, `ref_aa` and `ref_seq_window`, include two trivial baselines, and a weak sequence floor would make it easy for no sequence-only cell to fall below its floor. The six sequence floors of the comparison therefore also include a profile floor, `ref_seq_expert` (the ± 8 one-hot window under seven published per-residue scales, those scales smoothed over windows of 3, 7 and 15 residues, the chain's composition and the residue's position; 522 features), and a non-local floor, `ref_seq_global` (the same plus the 400 dipeptide frequencies and two long-range averages of the scales; 937 features), both defined in Table 41 with nothing fitted on a label. The profile floor improves on the window floor on every probe (for example RSA $0.526 \rightarrow 0.571$ and DSSP-3 $0.665 \rightarrow 0.689$). A third floor, `ref_seq_align_1nn`, transfers the label of the best MMseqs2 hit among the training domains of the same split and counts a position with no aligned counterpart as a failure. The split removes the homology it needs (median identity of the hit 0.357, about a quarter of positions aligned), so it reaches only 0.145 on DSSP-3 (0.494 on the positions it aligns). With these floors, 0 of 110 sequence-only cells fall below the sequence floor and 233 of 245 coordinate cells below theirs. The floors narrow the margins: the smallest, on CA distance bins, is $+0.003$ with all six sequence floors (Appendix B.2) against $+0.019$ with the four simplest, so there the sequence-only encoders and a fixed function of the sequence are indistinguishable.

C CONTROLS FOR THE DOWNSTREAM COMPARISON

This section holds the controls behind the downstream comparison of Section 3.2: a reference that uses no learned representation, the same checkpoint with and without its structure input, and the group comparison with pretraining, scale, the training budget and the input width each held fixed.

C.1 A NON-LEARNED REFERENCE FOR THE DOWNSTREAM TASKS

How much of each downstream score does a non-learned description of the sequence already reach? The answer is the basis of the main-text statement that the median encoder clears a composition reference on all ten tasks, by less than 0.02 on four of them (Section 3.2). The probing floors ask the same question of the probes. The reference, `ref_composition`, is the 20 amino-acid frequencies, $\log L$ and $L/600$, entered into the pipeline as an encoder and scored through the identical protocol: the same standardisation, linear and MLP-2 heads, validation-based selection and three seeds (`scripts/make_downstream_reference.py`).

Table 13: Every downstream task against a reference that uses no learned representation. The reference is amino-acid composition, $\log L$ and $L/600$ under the identical protocol.

task	metric	composition reference	median encoder	best encoder	encoders above
EC	accuracy	0.086	0.385	0.819	31 / 33
GO-MF	F-max	0.227	0.290	0.565	29 / 33
GO-BP	F-max	0.274	0.285	0.405	24 / 33
localisation	accuracy	0.470	0.485	0.705	21 / 33
homology	accuracy	0.071	0.151	0.403	23 / 32
solubility	AUROC	0.588	0.598	0.707	19 / 33
membrane	AUROC	0.779	0.826	0.913	26 / 33
stability	Spearman	0.090	0.284	0.777	29 / 33
fluorescence	Spearman	0.362	0.378	0.587	22 / 33
affinity	Spearman	0.091	0.249	0.499	30 / 33

The median encoder clears the reference on all ten tasks. The margins differ by task: EC $+0.300$ and stability $+0.193$ for the median encoder, against $+0.011$ on GO-BP, $+0.010$ on solubility, $+0.014$ on localisation and $+0.016$ on fluorescence. On those four tasks the lower half of the panel is within 0.02 of a 23-dimensional descriptor of the sequence, while the best encoder clears it by 0.119 to 0.235 (Table 13) and the median of the eight pretrained sequence models of at least 100M parameters

by 0.103 (solubility) to 0.205 (localisation). The closeness to the median encoder therefore comes from the small and randomly initialised encoders of the panel. The repair test of Section 3.3 asks whether the tasks use composition. On these four tasks it names composition on fluorescence, finds a smaller repair that holds across architecture families on solubility, and finds none that holds across families on GO-BP or subcellular localisation (Appendix D.11).

C.2 Removing the structure input from the same checkpoint

The group comparisons of Section 3.2 set encoders with different inputs against each other, and those encoders also differ in architecture, pretraining data and size. This subsection holds the weights fixed and changes only the input. ESM3 is run with the true sequence and no structure (no coordinates, so no structure tokens and no input to its geometric attention), and SaProt with every 3Di token replaced by the symbol #, which its vocabulary provides for a residue without a structure token. Both sequence-only sources pass the input audit as not reading coordinates (coordinate swap, jitter and rigid motion change the embedding by 0.0) and as reading residue identity (a residue shuffle changes it by a relative 0.28 to 0.60). They are scored with the released protocol on the ten tasks and the fifteen probes of the floor comparison. The released ESM3 and SaProt readouts are refitted and reproduce the released scores exactly, and each difference carries a paired bootstrap interval over test proteins (2,000 draws). The specification was fixed before scoring (`extract/ab_modality/PRESPEC.md`), and four SaProt embedding sets computed on a second machine agree with a local recomputation to a relative 3.3×10^{-7}.

Downstream tasks (Table 14). The structure input raises fold recognition for both models (ESM3 $+0.055$, SaProt $+0.117$) and stability for ESM3 ($+0.074$). For SaProt it also raises EC ($+0.252$), GO-MF ($+0.086$), GO-BP ($+0.041$) and PPI affinity ($+0.258$); for ESM3 the three annotation tasks do not change ($|\Delta| \leq 0.003$). It lowers fluorescence for both models (ESM3 -0.119, SaProt -0.136), a task whose ESMFold structures have a median pLDDT of 46.9, and solubility and localisation for both (-0.016 to -0.035); for ESM3 it also lowers PPI affinity (-0.078) and membrane versus soluble (-0.008), and for SaProt stability (-0.025). Counting intervals, the structure input helps on 2 of the ten tasks for ESM3 and on 5 for SaProt, and costs on 5 and 4. On the four tasks with experimental structures (Table 45) it costs on none; on the six with predicted structures it costs on five for ESM3 and four for SaProt and helps on one each (stability for ESM3, PPI affinity for SaProt).

Is # a valid absence of structure? Replacing every structure token by # could be an input SaProt handles badly for reasons unrelated to structure. A third source keeps each protein's own 3Di tokens and permutes their positions (a fixed seed from the protein identifier), so the 3Di composition is kept and the correspondence between residue and structure is lost (`extract/ab_modality/PRESPEC_shuf.md`). The full model scores above the shuffled one on all ten tasks ($+0.069$ to $+0.458$) and all fifteen probes, and the shuffled model scores below the # model on eight tasks and twelve probes (Table 16). Misaligned structure tokens therefore cost more than absent ones, and the # input behaves as missing structure.

Structural probes (Table 15). Removing the structure input lowers the probe score on 29 of the 30 cells, by 0.067 to 0.262 on the eleven labels the stored atoms determine; ESM3 on fold architecture is unresolved. Read against the floor of the input each version receives, the verdict follows the input. With the structure input, SaProt is below the floor of its atoms on all eleven determined labels and ESM3 on eight, the other three being the ideal-geometry cells of Appendix B.3. With the same weights and the sequence only, both are above the sequence floors on all eleven. On the four undetermined labels all four sources are above their floors. The same weights move from below the floor of their input to above it when only the input changes, which is the pattern the determinacy reading of Section 3.1 predicts.

C.3 The group comparison with each confound held fixed

Does the sequence-only lead on the downstream tasks come from the input, or from pretraining and scale? This subsection gives the tests behind the main-text statement that among large pretrained encoders the two groups are not separated (Section 3.2). The sequence group is ten pretrained language models from 8M to 3B parameters. The coordinate side contains ten randomly initialised networks and several below a million parameters. A gap between the two groups therefore mixes input modality with pretraining and scale. The panel contains no large randomly initialised sequence model and no

Table 14: Within-model modality ablation on the downstream tasks. S is the released protocol (linear or MLP-2 readout chosen on validation, mean of three seeds). $\Delta = S(\text{full}) - S(\text{sequence only})$ with a 95% paired bootstrap interval over test proteins (test pairs for PPI affinity), 2,000 draws. $^{+}$: interval above zero; $^{-}$: interval below zero.

		ESM3			SaProt		
Task	Metric	full	seq. only	Δ [95% CI]	full	seq. only	Δ [95% CI]
Experimental structures (PDB)							
EC number	Acc.	0.802	0.799	+0.003 [-0.004, +0.009]	0.687	0.435	+0.252 [+0.235, +0.268]$^{+}$
GO-MF	$F_{\max}$	0.533	0.531	+0.002 [-0.002, +0.006]	0.486	0.400	+0.086 [+0.078, +0.094]$^{+}$
GO-BP	$F_{\max}$	0.381	0.384	-0.003 [-0.005, +0.001]	0.369	0.328	+0.041 [+0.037, +0.045]$^{+}$
Remote homology	Acc.	0.403	0.347	+0.055 [+0.030, +0.082]$^{+}$	0.298	0.181	+0.117 [+0.082, +0.150]$^{+}$
Predicted structures (AlphaFold DB)							
Localisation	Acc.	0.616	0.638	-0.021 [-0.032, -0.010]$^{-}$	0.659	0.675	-0.016 [-0.029, -0.003]$^{-}$
Membrane vs soluble	AUROC	0.889	0.896	-0.008 [-0.012, -0.003]$^{-}$	0.902	0.901	+0.001 [-0.006, +0.009]
Predicted structures (ESMFold)							
Stability	Spearman	0.777	0.703	+0.074 [+0.058, +0.090]$^{+}$	0.715	0.740	-0.025 [-0.040, -0.010]$^{-}$
Fluorescence	Spearman	0.459	0.579	-0.119 [-0.143, -0.096]$^{-}$	0.427	0.563	-0.136 [-0.162, -0.109]$^{-}$
Solubility	AUROC	0.705	0.732	-0.027 [-0.048, -0.005]$^{-}$	0.680	0.715	-0.035 [-0.061, -0.012]$^{-}$
PPI affinity	Spearman	0.338	0.416	-0.078 [-0.106, -0.046]$^{-}$	0.439	0.182	+0.258 [+0.201, +0.313]$^{+}$

Table 15: Within-model modality ablation on the structural probes (superfamily split, symmetric protocol). $\Delta = S(\text{full}) - S(\text{sequence only})$ with a 95% protein-level paired bootstrap interval (2,000 draws); readout family chosen on validation for each source. Floor: the sequence-only source against its input-matched floor (best of the six sequence floors, chosen on validation); ↑ above, ↓ below, ∘ unresolved at 95%.

	ESM3		SaProt	
Probe	Δ [95% CI]	seq. only vs floor	Δ [95% CI]	seq. only vs floor
Determined by the backbone				
ϕ (circular)	+0.074 [+0.070, +0.077]$^{+}$	+0.098 ↑	+0.078 [+0.075, +0.081]$^{+}$	+0.048 ↑
ψ (circular)	+0.147 [+0.139, +0.155]$^{+}$	+0.334 ↑	+0.246 [+0.237, +0.256]$^{+}$	+0.182 ↑
Local curvature	+0.134 [+0.128, +0.141]$^{+}$	+0.336 ↑	+0.229 [+0.221, +0.237]$^{+}$	+0.195 ↑
DSSP-3	+0.074 [+0.069, +0.079]$^{+}$	+0.206 ↑	+0.136 [+0.131, +0.142]$^{+}$	+0.113 ↑
DSSP-8	+0.213 [+0.201, +0.228]$^{+}$	+0.274 ↑	+0.262 [+0.249, +0.274]$^{+}$	+0.112 ↑
Core vs surface	+0.067 [+0.062, +0.071]$^{+}$	+0.181 ↑	+0.125 [+0.119, +0.130]$^{+}$	+0.091 ↑
Contact density	+0.161 [+0.153, +0.169]$^{+}$	+0.205 ↑	+0.207 [+0.201, +0.214]$^{+}$	+0.079 ↑
Distance bins	+0.102 [+0.097, +0.107]$^{+}$	+0.114 ↑	+0.127 [+0.123, +0.131]$^{+}$	+0.044 ↑
Contact order	+0.099 [+0.085, +0.115]$^{+}$	+0.401 ↑	+0.151 [+0.133, +0.169]$^{+}$	+0.327 ↑
Radius of gyration	+0.154 [+0.114, +0.202]$^{+}$	+0.274 ↑	+0.142 [+0.094, +0.207]$^{+}$	+0.257 ↑
Asphericity	+0.112 [+0.089, +0.137]$^{+}$	+0.289 ↑	+0.121 [+0.096, +0.149]$^{+}$	+0.232 ↑
Not determined by the backbone				
RSA	+0.126 [+0.120, +0.132]$^{+}$	+0.269 ↑	+0.207 [+0.200, +0.213]$^{+}$	+0.137 ↑
RSA bins	+0.134 [+0.128, +0.140]$^{+}$	+0.204 ↑	+0.178 [+0.173, +0.183]$^{+}$	+0.090 ↑
B-factor	+0.070 [+0.064, +0.077]$^{+}$	+0.153 ↑	+0.119 [+0.110, +0.128]$^{+}$	+0.069 ↑
CATH architecture	-0.039 [-0.075, +0.048]	+0.362 ↑	+0.056 [+0.013, +0.158]$^{+}$	+0.299 ↑

coordinate model above 1.5B, so it cannot separate these confounds, and we report each restriction separately. The 33 encoders are also related: ESM-2 appears at five scales, PottsMPNN four times, and five architectures appear in a random and a pretrained version (GCPNet, GearNet, IPA, PottsMPNN and TFN; the random GearNet is a smaller configuration, 7M against 20M parameters). Every test is therefore also run on one value per architecture family (`scripts/group_controls.py`). The grouping is the measured input audit, so ProstT5, declared multimodal and measured not to read coordinates, is on the sequence side.

The lead shrinks as the restriction tightens (Table 17). Over all encoders the sequence group leads on all ten tasks by a mean of $+0.186$, and the family test is significant on eight. Dropping the randomly initialised networks leaves the lead on all ten at $+0.162$, significant on seven. Restricting both sides to pretrained encoders of at least 100M parameters (eight sequence models against four coordinate models) leaves a mean gap of $+0.048$, a lead on eight of the ten tasks, and significance on three.

Table 16: Shuffled-3Di control for SaProt on the ten downstream tasks (left) and the fifteen probes (right; below the rule, the four labels the backbone does not determine). Full: the real 3Di token at each residue; shuf.: the protein's own 3Di string permuted over positions (same letters and composition, wrong residue alignment); #: every 3Di token masked. Scores use the released protocol with the readout chosen on validation. Δ_A = full − shuf., Δ_B = shuf. − #; $^+$ / $^-$: the 95% paired bootstrap interval over test proteins (2,000 draws) lies above / below zero. The intervals are released (`extract/ab_modality/table_shuf3di.tex`, `table_shuf3di_probing.tex`).

Task	full	shuf.	#	Δ_A	Δ_B
EC number	0.687	0.277	0.435	+0.410$^+$	-0.158$^-$
GO-MF	0.486	0.278	0.400	+0.208$^+$	-0.122$^-$
GO-BP	0.369	0.285	0.328	+0.084$^+$	-0.043$^-$
Remote homology	0.298	0.160	0.181	+0.138$^+$	-0.021
Localisation	0.659	0.512	0.675	+0.147$^+$	-0.163$^-$
Membrane vs soluble	0.902	0.834	0.901	+0.069$^+$	-0.067$^-$
Stability	0.715	0.257	0.740	+0.458$^+$	-0.483$^-$
Fluorescence	0.427	0.205	0.563	+0.222$^+$	-0.358$^-$
Solubility	0.680	0.608	0.715	+0.072$^+$	-0.107$^-$
PPI affinity	0.439	0.145	0.182	+0.295$^+$	-0.037

Probe	full	shuf.	#	Δ_A	Δ_B
ϕ (circular)	0.943	0.820	0.865	+0.123$^+$	-0.045$^-$
ψ (circular)	0.935	0.526	0.688	+0.409$^+$	-0.162$^-$
Local curvature	0.943	0.550	0.714	+0.393$^+$	-0.164$^-$
DSSP-3	0.919	0.676	0.782	+0.243$^+$	-0.107$^-$
DSSP-8	0.670	0.292	0.408	+0.378$^+$	-0.116$^-$
Core vs surface	0.886	0.671	0.761	+0.215$^+$	-0.090$^-$
Contact density	0.652	0.367	0.445	+0.285$^+$	-0.078$^-$
Distance bins	0.294	0.114	0.166	+0.179$^+$	-0.052$^-$
Contact order	0.901	0.737	0.750	+0.164$^+$	-0.013
Radius of gyration	0.880	0.809	0.738	+0.071$^+$	+0.071$^+$
Asphericity	0.801	0.667	0.680	+0.134$^+$	-0.013
RSA	0.914	0.561	0.708	+0.353$^+$	-0.146$^-$
RSA bins	0.801	0.523	0.623	+0.278$^+$	-0.100$^-$
B-factor	0.612	0.393	0.493	+0.219$^+$	-0.100$^-$
CATH architecture	0.465	0.263	0.409	+0.202$^+$	-0.146$^-$

Table 17: The sequence-versus-coordinate comparison under each restriction, at each encoder's native width and with every representation projected to 116 dimensions. The family test is a Mann–Whitney test over architecture families, one value per family. Holm: tasks whose family-test p stays below 0.05 after Holm correction over the ten tasks. Mean rank gap: the mean over tasks of the coordinate families' mean rank minus the sequence families' (ranks within the 33-encoder panel, positive when the sequence side ranks higher), with a permutation p over family labels (10,000 permutations). Normalised gap: per task, scores rescaled so that the composition reference is 0 and the best encoder 1.

			tasks the sequence group leads		mean gap over tasks		tasks, family test $p < 0.05$	
restriction	encoders	families	native	116 dims	native	116 dims	native	116 dims
all encoders	10 / 23	6 / 15	10 of 10	10 of 10	+0.186	+0.164	8 of 10	9 of 10
pretrained only	10 / 13	6 / 10	10 of 10	10 of 10	+0.162	+0.155	7 of 10	9 of 10
pretrained, at least 100M parameters	8 / 4	6 / 4	8 of 10	9 of 10	+0.048	+0.055	3 of 10	4 of 10

restriction	tasks after Holm	mean rank gap (permutation p)	normalised gap
all encoders	8 of 10	+13.3 (< 0.001)	+0.619
pretrained only	7 of 10	+12.6 (0.006)	+0.561
pretrained, at least 100M parameters	0 of 10	+4.1 (0.093)	+0.196

Remote homology reverses: the coordinate side leads by 0.090. PPI affinity, the column Section 3.2 draws no conclusion from, also reverses, by 0.020.

Ten tasks tested one at a time need a correction for multiplicity, and a mean of raw gaps mixes four metrics (`scripts/group_controls_multiplicity.py`). Over all encoders and over the pretrained ones, the lead holds under Holm correction over the ten tasks (8 and 7 of 10) and under a family-level permutation test of the mean rank gap ($p < 0.001$ and $p = 0.006$). Among pretrained encoders of at least 100M parameters no task holds under Holm correction: the sequence families rank 4.1 places higher on average, with a permutation p of 0.093, and the normalised gap is +0.196 of the distance from the composition reference to the best encoder. The design cannot resolve the two groups at this scale. With six sequence and four coordinate families there are 210 assignments of families to groups, so the smallest two-sided p a family-level Mann–Whitney test can reach is $2/210 = 0.0095$, above the first Holm threshold over ten tasks (0.005); no task could pass whatever the data. A bootstrap over families puts the mean gap over the ten tasks at $+0.049$ $[-0.016, +0.114]$. Two of the four coordinate encoders in this restriction, ESM3 and SaProt, also read the sequence at scale, so the restriction does not match pretraining between the groups. Among large pretrained encoders the point estimates favour the sequence-only group and the panel cannot separate the two. The data therefore show no gain from reading coordinates for a pooled readout at this scale; they do not show that a sequence model is better, and pretraining is not matched between the groups.

The training budget. The downstream heads are trained on at most 6,000 proteins on eight of the ten tasks (Section 3.2). A shallow head on frozen embeddings at that budget could favour representations with better-conditioned global features, which would be the sequence group. Without re-extracting every encoder on the full splits the budget can only be lowered, so the same protocol is run at 25% and 50% of the training rows (`scripts/train_size_sweep.py`). The gap grows with the budget (Table 18): EC from $+0.359$ to $+0.473$, GO-MF from $+0.004$ to $+0.234$, stability from $+0.281$ to $+0.355$. The two tasks whose sign changes, GO-BP and PPI affinity, are the two with the smallest gaps at full budget. Over the range we can vary, a smaller budget shrinks the sequence-only lead, so the cap does not produce it. Budgets above the cap are untested.

Table 18: The group gap as the training budget is reduced. Positive means the sequence group leads.

task	25% of training rows	50%	100%
EC	+0.359	+0.458	+0.473
GO-BP	-0.017	+0.007	+0.090
GO-MF	+0.004	+0.106	+0.234
affinity	-0.031	+0.026	+0.011
fluorescence	+0.142	+0.160	+0.183
homology	+0.113	+0.119	+0.120
localisation	+0.190	+0.186	+0.189
membrane	+0.110	+0.103	+0.104
solubility	+0.079	+0.105	+0.105
stability	+0.281	+0.312	+0.355

Every representation at one width. The two heads of the downstream protocol are the same for every encoder (Appendix H.5), and their input runs from 20 to 2,560 dimensions. The sequence-only encoders are the wider group, and a wider input gives each head's first layer more parameters. To remove that difference, every representation wider than 116 dimensions is projected onto its top 116 principal directions, fitted on the training split alone, and scored again through the identical heads, selection rule, seeds and metrics (`extract/score2_fixed_dim.py`, `scripts/fixed_dim_report.py`). 116 is the native width of GVP and GCPNet, the narrowest encoders after Foldseek, which stays at its 20 dimensions, and the width of the matched-capacity probing control of Appendix B.7. The lead holds at a common width (Table 17). Over all encoders it is $+0.164$ at 116 dimensions against $+0.186$ at native width, with the sequence group ahead on all ten tasks in both. Among pretrained encoders of at least 100M parameters it is $+0.055$ against $+0.048$, ahead on nine of ten tasks against eight. Remote homology reverses at both widths (-0.094 and -0.090). PPI affinity reverses at native width (-0.020) and leads at 116 dimensions ($+0.105$); over all encoders it moves from a tie ($+0.011$) to a $+0.207$ lead. It is the task whose seed noise Section 3.2 sets aside, and the only one whose sign depends on the width.

D Controls for the erasure analysis

D.1 A positive control: can the protocol find a demand that was planted?

Can the erasure protocol find a demand whose answer is known in advance, and does it name nothing when there is nothing to find? The main text relies on this control for the statement that a planted demand is recovered in all three conditions and that with nothing planted every effect lies within ± 0.007 (Section 3.3). We build tasks whose answer we set. On the stability proteins and twelve encoders spanning the three classes, a synthetic label is generated as a fixed linear function of one property block plus noise at equal variance. The property used is the conditional residual the protocol erases (the block with the other independent blocks projected out), because the conditional arm is defined on that quantity. A fourth condition plants nothing: the label is built from a direction of the representation with every property block projected out, so any property the rule names there is a false positive. Everything downstream follows the path of Section 3.3 and the standard effect rule: the same LEACE erasure, the same sham null over three permutation seeds, the same sign test over encoders with Holm correction and the same magnitude floor (`scripts/erasure_positive_control.py`).

Table 19: Mean conditional effect when the label is a known function of one property, over the encoders whose erasure verified (twelve unless stated), which is the set the rule is computed on. Top: the planted label is linear in the property; bottom: it is a nonlinear function of the property. Bold marks what the standard effect rule names. The last column of each part is the condition in which nothing was planted.

erased property	planted amino-acid composition	planted secondary structure content	planted global shape	planted none
amino-acid composition	**+0.612**	-0.008	+0.020	-0.002
burial and packing	+0.005	+0.006	**+0.093**	+0.001
contact order	+0.003	-0.013	-0.003	+0.001
global shape	+0.000 ($n = 8$)	+0.003 ($n = 8$)	**+0.558** ($n = 8$)	+0.001 ($n = 8$)
secondary structure content	-0.002	**+0.200**	-0.025	+0.001
size	-0.007 ($n = 4$)	+0.006 ($n = 4$)	-0.053 ($n = 4$)	+0.007 ($n = 4$)

erased property	planted amino-acid composition, nonlinearly	planted global shape, nonlinearly	planted nothing
amino-acid composition	**+0.250**	+0.063	-0.002
burial and packing	-0.005	-0.054	+0.001
contact order	+0.014	+0.001	+0.001
global shape	-0.046 ($n = 8$)	-0.040 ($n = 8$)	+0.001 ($n = 8$)
secondary structure content	+0.023	-0.008	+0.001
size	+0.006 ($n = 4$)	+0.070 ($n = 4$)	+0.007 ($n = 4$)

The protocol recovers the planted property in all three conditions: $+0.612$ for amino-acid composition, $+0.558$ for global shape and $+0.200$ for secondary-structure content, against at most $+0.093$ for any other property in the same condition. It produces one false positive, burial and packing under planted global shape, which is the pair of blocks with the largest residual correlation after conditioning. In the condition where nothing was planted every effect lies between -0.002 and $+0.007$, the latter on size over the four encoders whose erasure verified, and the rule names nothing, so the false-positive rate of the rule on a task with no demand is 0 of the six properties it evaluates.

The control also tests the erasure check, which drops the cells whose erasure did not verify. Without the check, over all cells, the rule names nothing under planted composition or under planted secondary structure, because the cells where the erasure did not take reduce the sign test to noise. With the check the recovery is visible, at a cost in coverage: under planted global shape it keeps 8 of the 12 encoders.

A demand the task uses nonlinearly. If a task uses a property only through a nonlinear function of it, does the linear erasure still detect the demand? A second positive control measures this: the same construction with the synthetic label a nonlinear function of the planted property, run through the identical path over the same twelve encoders (Table 19, bottom). The standard effect rule still names amino-acid composition, at $+0.250$ against $+0.612$ for the linear plant. It names nothing when global shape is planted (-0.040 against $+0.558$), and nothing when nothing is planted, with every effect within ± 0.007. Without the erasure check it names nothing in any condition. A demand that a task uses only through a nonlinear readout of the property is therefore under-detected or missed, which is the direction the recoverability numbers of Appendix D.2 predict, and the demands of Section 3.3 are lower bounds on what the tasks need from these properties.

D.2 What the erasure removes: linear access

Does the erasure remove the property, or only linear access to it? The answer defines the main-text term linear-head demand (Section 2.4) and supports the statement that after erasure a two-layer MLP keeps a median 0.86 of its pre-erasure readout (Section 3.3). The erasure check in the sweep is a linear probe, and LEACE removes linear predictability by construction, so that check verifies that the map was applied. Whether the property is gone needs a readout LEACE does not constrain. We repeat the before-and-after measurement with a two-layer MLP (hidden width 256, 40 epochs) over four tasks (stability, solubility, membrane and soluble, subcellular localisation), eight encoders spanning the panel (ESM-2 650M, ESM-1b, ProtT5-XL, SaProt, ESM-IF1, ProteinMPNN, GVP, GearNet-pretrained) and the eight properties, 256 cells in all (`scripts/recoverability_check.py`, `erasure/recoverability_check.csv`).

The linear readout of the erased property falls to noise in every cell (Table 20; median $0.669 \rightarrow -0.003$ over the six independent properties; kept fraction 0.00). The MLP readout keeps a median

Table 20: Readout of the erased property before and after its LEACE erasure, median over the 32 cells of each property. The linear readout is what the sweep's erasure check measures; the MLP readout is what LEACE does not constrain. Charge and hydrophobic fraction are exact linear functions of amino-acid composition and are not readable before erasure either.

erased property	linear, before	linear, after	MLP, before	MLP, after	cells
amino-acid composition	0.902	-0.001	0.384	0.216	32
burial and packing	0.722	-0.060	0.672	0.589	32
charge	0.029	0.004	0.004	-0.009	32
contact order	0.653	-0.023	0.241	0.188	32
global shape	0.563	0.026	0.501	0.449	32
hydrophobic fraction	0.020	0.002	-0.007	0.008	32
secondary structure content	0.732	0.003	0.624	0.534	32
size	0.621	-0.022	0.584	0.537	32

0.86 of its pre-erasure value ($0.503 \rightarrow 0.352$), from 0.56 for amino-acid composition to 0.92 for burial and packing and 0.95 for size, between 0.81 and 0.93 on every task, and between 0.49 (GVP) and 0.96 (ESM-IF1) across encoders. The erasure therefore removes linear access to the property and leaves most of the property in place. Every effect in Section 3.3 is the cost to the downstream head of losing linear access, and the demands named there are linear-head demands, demands for linearly accessible information.

D.3 The eraser against the reference LEACE implementation

Section 3.3 erases a property with the closed-form LEACE projection. This subsection asks whether the eraser used throughout (`scripts/erasure.py`) matches the reference LEACE implementation (the `concept-erasure` package), and whether the named fluorescence pairs depend on where the two differ. We compared them on 64 synthetic configurations and on the embeddings of three encoders (`protcompass/qd/leace_check/`). On synthetic data the two agree to 0.17% in double precision at matching settings, and after our map a ridge probe recovers none of the concept on held-out data. On real embeddings our map removes the linear cross-covariance with the property, as LEACE does. Our map also applies an absolute floor of 10^{-6} to the eigenvalues of the representation's covariance when it whitens, which treats the smallest directions of wide encoders as null, so it removes more of the representation than the minimal edit of exact LEACE (2.1% against 0.8% of ESM-2 650M's variance for burial and packing, 13.0% against 6.9% of ProteinMPNN's for composition). The erased representation also comes in an arbitrary basis to which the standardising head is sensitive, and the basis moves ESM-2 650M's repair on burial and packing between $+0.005$ and $+0.035$; encoders of about 120 dimensions move by 0.002. Both fluorescence pairs pass the repair rule under all five variants we ran: the map used throughout, a principal-axes basis, a rotation-invariant head, and exact LEACE with either head (composition $+0.025$ to $+0.028$, burial and packing $+0.021$ to $+0.023$). The fluorescence verdicts are therefore the same under both erasers.

The rank cutoff. Separately from that eigenvalue floor, the eraser cuts the rank of the whitened cross-covariance at 10^{-6} of its leading singular value, and every runner in the release uses it, `scripts/bighead_holdout.py` included. The cutoff matters on the four multi-output tasks. A cutoff of 10^{-8} keeps one numerically null direction there, and erasing a null direction with an oblique projection shears the representation: on EC classification and ESM2-650M the erased arm then has 167 times the variance of the unerased one and the sham arm 298 times. With the shared eraser the rank on that cell is 19 and the erased arm removes $+0.159$ of the variance, the numbers Appendix E.4 reports.

D.4 The null the erasure is read against

The sham applies the same LEACE map to a permuted target. A random subspace of the same rank is an invalid null for this erasure. LEACE's un-whitening step biases the removed subspace toward high variance, so the erasure removes 6.7 to 23 times more of the representation's variance than the rank-matched control, and under that control a property row-permuted across proteins, which carries no relation to the protein it is assigned to, passes a sign test over encoders at $p = 0.021$. We

therefore use as the sham a LEACE erasure of the same property permuted across proteins. Alongside it we report a top-r principal-direction erasure as the variance ceiling, verify after every erasure that the property is no longer linearly recoverable, and require that a property vary on a task before a demand for it can be claimed. Two of the eight properties, hydrophobic fraction and charge, are exact linear functions of amino-acid composition, so their conditional residual is float32 rounding error; they are measured as contrasts inside composition and labelled as such.

A random rotation is not a null. Against a random-rotation control the mean effect is higher by $+0.003$ than against the sham, 8% of cells change sign between the two controls, and the rotation arm scores *above* the unerased representation in 49% of cells, which a null should not do. Under the rotation control amino-acid composition leads every task. Under the sham it leads five of six, and on fluorescence, the one task where a geometric property leads a compositional one, burial and packing leads.

The sham removes less than the erasure. The sham and the erased arm match in rank and in the change of basis, and they differ in how much of the representation they remove: permuting destroys the relation between the target and the representation, so the fit finds little to take. Over 1,225 audited cells of the released sweep (the ten tasks against amino-acid composition, burial and packing, contact order and secondary-structure content, with 31 to 33 encoders per pair on nine tasks and 11 to 13 on PPI affinity), the real erasure removes a median 3.6 times more of the representation's variance than the sham (interquartile range 0.9 to 8.3), and more in 895 of them. On amino-acid composition the two average 0.361 and 0.159, and the erasure removes a median 0.565 of the property's linear recoverability where the sham removes almost none (`scripts/sham_variance_audit.py`). The effect therefore charges two things to the property at once: the directions the map removed and the amount that went with them. The sham keeps nearly the whole representation while taking the change of basis, which is why it can score above the unerased representation where the head is overparameterised. In the whitened geometry the erasure is defined in, any rank-r removal takes the same amount; the asymmetry appears in the original coordinates the head reads.

A null matched on amount. `scripts/matched_null.py` adds a second null that matches the rank, the change of basis and the amount removed. It removes a rank-r subspace spanned by principal directions of the training representation whose total variance equals the real map's to within 1%, chosen among the subsets meeting that constraint to carry as little of the property as a randomised swap search finds; principal directions are used because their scores are uncorrelated, so the property variance a subset removes is the sum of its members'. If the erased arm still costs more than this null, the extra cost is specific to the property and is not the capacity removed. Matching the amount and avoiding the property are compatible: on ESM2-650M and membrane vs soluble a rank-19 subspace removing 13.3% of the variance carries 6% of the property, against 95% for the real erasure, and over the audit the null carries a median 0.010 of the property against the erasure's 0.565. The variance target is met to within 1% on 754 of the 1,225 audited cells (62%) and within 5% on 1,026 (84%). The misses concentrate where the erasure removes little, mostly properties of rank one to three, because the subspace must be a whole number of principal directions. On the 60-pair grid of Appendices D.6 and D.11, six properties on the ten tasks and 1,964 cells, the null is matched within 1% on 1,245 cells (63.4%). For composition on fluorescence, secondary structure on PPI affinity and burial and packing on fluorescence the shares are 84%, 100% and 73%, and on its matched cells burial and packing keeps $R_{\mathrm{var}} = +0.018$ on 24 of 24 encoders. For GCPNet the match misses by 20 to 35%. Matching the fitting procedure instead, with the amount left free, fails as the sham does: a surrogate target of the same width whose linear recoverability equals the property's removes 0.05% of the variance against the real map's 13%, because the directions free of the property are the low-variance tail.

What the matched null changes. Table 21 reruns the conditional arm of the released sweep with the matched null added, on 32 or 33 encoders per pair, and gives the effect behind each named pair.

Erasing the linearly accessible subspace of amino-acid composition removes about a third of these representations' variance (0.361 on average), and a comparison of levels cannot separate a loss of that size from the capacity removed: against the matched null only burial and packing on fluorescence and composition on GO-MF keep their sign across encoders, and on seven of the nine rows Δ_{var} does not. A capacity loss common to both arms cancels in a difference of differences and remains in

task	property	n	Δ_{sham}		Δ_{var}	
			mean	positive	mean	positive
fluorescence*	burial and packing	33	+0.0534	33	+0.0338	29
stability	amino-acid composition	32	+0.0658	23	+0.0280	21
GO-MF*	amino-acid composition	32	+0.0204	29	+0.0256	30
fluorescence*	amino-acid composition	32	+0.0392	30	+0.0090	22
PPI affinity*	secondary structure content	33	+0.0685	33	+0.0085	20
solubility*	amino-acid composition	32	+0.0235	32	+0.0062	19
membrane vs soluble*	amino-acid composition	32	+0.0501	32	−0.0040	13
subcellular localisation*	amino-acid composition	32	+0.0353	32	−0.0408	12
PPI affinity*	contact order	33	+0.0385	28	−0.0574	15

Table 21: The released conditional effect of every pair the standard effect rule names (*), and of stability composition, against the sham and against the variance-matched null, ordered by the second. Against the variance-matched null a two-sided sign test over encoders with Holm correction over the nine rows is passed by burial and packing on fluorescence and composition on GO-MF and by nothing else (`scripts/erasure_v2.py --matched-null`, `scripts/matched_null_stage1_report.py`).

a difference of levels, so the effect is reported bracketed by its two nulls and the naming rests on the repair contrast (Appendix D.6). The held-out cohorts show the same split: against the matched null the effect loses its sign consistency and the repair contrast keeps it (Appendix E.4).

Two runs. Table 21 and Table 23 give Δ_{var} for these nine pairs from two separate runs, each of which fits its own erasures, shams, nulls and heads. Under a two-sided sign test with Holm correction over the nine rows both runs keep the same two pairs and lose the other six of the eight the standard effect rule names, so the statement of Section 3.3 holds in both. The means agree to within 0.0021 on six of the nine rows and differ most on PPI affinity, whose test split has 326 pairs and a mean seed standard deviation of 0.101 against 0.011 on the other nine tasks. The grid cells of fluorescence and PPI affinity were computed on a second machine with different numerical libraries (NumPy 2.5.3, SciPy 1.18.1 and scikit-learn 1.9.1, against NumPy 1.26.4, SciPy 1.15.3 and scikit-learn 1.7.2), where the refitted erasure removes a different share of the variance in nearly every cell, by up to 0.34 in one; the code, embeddings, property values and splits are identical on the two machines.

D.5 What the effect rule can detect, and how its verdicts depend on the settings

On which tasks does the erasure detect a demand, on which does it bound one, and on which can it decide neither? Table 22 states a one-sided 95% upper confidence limit on each property's mean effect over encoders, Holm-adjusted over the six properties within a task, and calls a task *bounded* only when every limit falls below the magnitude floor $\delta = 0.02$ of the effect rule. An upper limit bounds the effect itself; a minimum detectable effect says only that the design would probably have missed an effect of that size. A bound is read against the reachable range drawn in Figure 4A: the loss from erasing the task label's own subspace, read at the largest rank a property block occupies on that task.

The standard effect rule names eight of the 60 pairs: amino-acid composition on membrane and soluble (+0.051), fluorescence (+0.036), subcellular localisation (+0.035), solubility (+0.024) and GO-MF (+0.020); burial and packing on fluorescence (+0.050); and, on PPI affinity, secondary-structure content (+0.067) and contact order (+0.040). Eight tasks are detections. Remote homology is the one bound: every upper limit is at or below +0.009, which is 15% of the +0.060 that erasing the label's own subspace costs at rank 19, so a demand of 0.02 or more is excluded there for the six properties erased. EC classification is an abstention: erasing the label's own subspace at the ranks the properties occupy costs nothing (−0.008), so there is no positive scale to express a bound against, the upper limits themselves reach +0.022, and EC is not drawn in Figure 4.

The rank ladder. The ladder erases the subspace LEACE fits to the task label itself, rank by rank, and retrains the head. On EC classification it is flat and slightly negative over eight encoders at every rank from 1 to 32, between −0.009 and −0.038 against a base of 0.523, and turns positive only

Table 22: Every task as a detection, a bound or an abstention. **Property rank** is the largest rank a property block occupies on that task; **erasing the label** is the reachable range read at or below that rank, with the rank it is attained at; **upper limit** is the largest Holm-adjusted one-sided 95% limit over the six properties, with its share of the reachable range.

task	encoders	property rank	erasing the label	largest effect	upper limit (Holm)	decision
homology	32	19	+0.060 at r=19	+0.005	+0.009 (15%)	bounded
fluorescence	33	20	+0.451 at r=1	+0.050	+0.056 (12%)	detected
GO-BP	33	19	+0.113 at r=19	+0.014	+0.017 (15%)	detected
GO-MF	33	19	+0.124 at r=19	+0.020	+0.026 (21%)	detected
membrane	33	19	+0.324 at r=1	+0.051	+0.058 (18%)	detected
affinity	33	40	+0.481 at r=1	+0.067	+0.082 (17%)	detected
solubility	33	19	+0.130 at r=1	+0.024	+0.027 (21%)	detected
stability	33	18	+0.433 at r=1	+0.070	+0.105 (24%)	detected
localisation	33	19	+0.204 at r=8	+0.035	+0.039 (19%)	detected
EC	33	19	-0.008 at r=1	+0.012	+0.022	inconclusive

above the ranks the property blocks occupy: $+0.042$ at rank 64, $+0.143$ at 128 and $+0.376$ at 256. Remote homology behaves the same way from a lower base: flat to rank 3, $+0.053$ at 19, $+0.131$ at 64 and $+0.228$ at 256. The 33-encoder erasure arm agrees at the ranks the property blocks occupy (EC -0.010). Because the ladder runs to rank 256 on these two tasks and to rank 1 on five others, the reachable range is always read at the rank of the property blocks.

What a label-directed erasure bounds. A flat ladder bounds the loss from erasing any other subspace of the same rank only for a whitened representation, squared loss, the optimal unregularised linear readout and the population covariance. With $\mathrm{Cov}(H) = I$, $C = \mathrm{Cov}(H, Y)$ and Q an orthogonal projector of rank r, the loss from the removal is then

$$\mathrm{risk}\big((I-Q)H\big) - \mathrm{risk}(H) \;=\; \|QC\|_F^2 \;\le\; \sum_{j=1}^{r} \sigma_j(C)^2, \tag{2}$$

with equality for the subspace the label selects, so under these conditions the label-directed removal is the most damaging one of its rank (simulated at $d = 12$, $k = 4$ and $n = 200{,}000$, the identity holds to a relative 5×10^{-6}). The scores we report meet none of the four conditions: the metrics are accuracy, F-max, AUROC and Spearman, the heads are retrained and nonlinear in 270 of the 329 cells, and the representations are not whitened. A label carried jointly by two directions shows the failure: with $y = \mathrm{sign}(h_1 h_2)$ over eight standard normal directions, erasing the rank-1 subspace the label selects costs a retrained two-layer head 0.032 in accuracy, while erasing h_1, also rank 1, costs 0.434 and leaves the head at chance. We therefore read the ladder as an empirical sensitivity curve for this erasure, this panel and these heads, and the exclusions as exclusions over the six properties erased, not over all subspaces of the rank they occupy. On EC classification the ladder still shows that the score is spread over many directions: the label's own subspace carries none of it up to rank 32 and $+0.376$ at rank 256.

The decision threshold. The magnitude floor of 0.02 is a reporting choice, set at roughly the largest sham spread we observe on any task, so that a demand has to exceed what the null itself moves. Lowering it to 0.01 adds three pairs and one task, GO-BP; raising it to 0.05 leaves three pairs on three tasks. The tasks that name a property at 0.02 are a subset of those that name one at 0.01 and a superset of those at 0.05, so the conclusion is monotone in the threshold.

The encoder subset. The sign test runs over all 33 encoders, and some of them score at the non-learned reference on a task (GearNet-pretrained reaches 0.036 on EC against a sequence-group mean of 0.745). Erasing a subspace from such a representation cannot produce a drop, so these cells push toward "no demand". Recomputed over the encoders that clear the composition reference of Table 13 on each task, the named set is unchanged on nine of the ten tasks. GO-MF loses its name: amino-acid composition reaches $+0.018$ over the 29 encoders that clear the reference, below the floor. It is the pair that also fails under family weighting and on the held-out cohort, and the paper does not name it (Appendices D.11 and E.4). The detection bounds move by at most 0.007, and remote

homology and EC keep their status, with largest effects of $+0.007$ and $+0.013$ against bounds of $+0.012$ and $+0.016$. The restriction uses the downstream scores to select encoders for an analysis of downstream scores, so it is a sensitivity analysis and the unrestricted numbers remain the headline.

D.6 THE REPAIR RULE AND THE PAIRS IT NAMES

Which task-property pairs does the repair rule name? This subsection gives the rule and the full grid behind the main-text statement that seven pairs pass (Section 3.3). Appendix D.4 shows that the erasure effect cannot be attributed to the property once the null removes as much of the representation as the erasure does, and that the repair contrast can. Section 3.3 therefore names pairs with a rule on the repair contrast, the repair rule, and the named set is derived from it here (`scripts/addback_named_pairs.py`, `scripts/named_pair_verdicts.py`).

Candidates. Every pair of the six non-redundant property blocks and the ten scored tasks, 60 pairs, is tested with no screening (`erasure/named_pair_grid_full.csv`). An earlier run screened to the 18 pairs whose effect against the sham clears a sign test; that screen rests on the quantity Appendix D.4 shows to be mostly capacity, so it can exclude a pair the repair rule would name.

Arms. For every pair and every encoder with an extraction, the arms are the erased representation, four shams and four variance-matched nulls, each scored with no descriptor, with the property's descriptor block and with that block's rows permuted across proteins. The head is the one the standard effect rule uses (`erasure.fit_score`), so the effect and the repair contrast come from one protocol.

Rule. The repair rule is the standard effect rule with its contrast replaced by R_{var} and nothing else changed: R_{var} must pass a two-sided sign test over encoders at Holm-corrected $p < 0.05$ over the 60 pairs, its mean over encoders must be at least 0.02, and the property must vary on the task. The 0.02 floor is on the score scale and was set with the effect rule, before any repair contrast was computed. The repair contrast estimates the share of the score the property accounts for, which is what the floor bounds, so the floor carries over without being chosen again. The sign test alone is reported as well, as presence without size. With 32 to 33 encoders it detects repairs of a few thousandths: on its own it passes 18 of the 60 pairs, two of them with a negative mean (contact order on fluorescence and on stability).

On EC classification the effect against the null is $+0.089$ to $+0.134$ for all six properties, positive on 32 or 33 of 33 encoders, while on the held-out proteins it is -0.046 for composition (Table 36). With 373 classes and a linear head every restricted arm scores above the unerased representation on this task (Appendix E.4), and erasing the label's own subspace costs nothing at these ranks (Appendix D.5), so these values reflect how the arms are conditioned; we do not read them as effects, and the repair rule names nothing on EC.

What the rule names. Seven pairs pass (Table 23). Four are on PPI affinity, secondary structure content ($+0.072$), contact order ($+0.056$), size ($+0.034$) and global shape ($+0.026$), on a split in which every test pair shares a chain with a training pair; Appendix D.13 tests them with chain-disjoint folds. The other three are burial and packing ($+0.022$, 33 of 33 encoders) and amino-acid composition ($+0.023$, 26 of 32) on fluorescence, and composition on stability ($+0.066$, 26 of 32). Stability composition is the pair the screen excluded (Holm $p = 0.15$ on the effect against the sham); on the held-out cohort it repairs on 12 of 15 encoders ($+0.045$), which is the decidable half of hypothesis D3 in Appendix E.4. Composition on solubility ($+0.019$) and on GO-MF ($+0.019$) passes the sign test over encoders with a mean just under the 0.02 floor, and every other property on those two tasks repairs by 0.008 or less. Two pairs the standard effect rule names fail the sign test: composition on membrane vs soluble ($+0.017$, 24 of 32, Holm $p = 0.27$; on the held-out cohort $+0.035$ on 15 of 15) and composition on subcellular localisation, where the variance-matched null costs the seven largest pretrained language models 0.11 to 0.32 more than the erasure itself, and composition repairs that larger loss by more than it repairs the erasure. The repair contrast assumes that the descriptor's gain does not depend on how much of the representation was removed, and the assumption fails when the two arms sit this far apart. With the logistic head of Appendix D.8 composition repairs on 5 of 6 encoders on the same cells, and on the held-out cohort on 12 of 15, so on this task the verdict depends on the readout.

			effect vs. null		repair contrast vs. null				one vote per family	
task	property	n	mean	positive	mean	positive	Holm p	named	positive	named
fluorescence*	burial and packing	33	$+0.0182$	27	$+0.0218$	33	1.4×10^{-8}	yes	21/21	yes
PPI affinity*	secondary structure	33	$+0.0188$	22	$+0.0721$	32	4.7×10^{-7}	yes	20/21	yes
solubility*	composition	32	$+0.0069$	19	$+0.0193$	31	8.9×10^{-7}	sign only	20/20	
GO-MF	burial and packing	33	$+0.0221$	30	$+0.0031$	29	6.2×10^{-4}	sign only	19/21	
GO-MF	size	33	$+0.0299$	31	$+0.0084$	29	6.2×10^{-4}	sign only	19/21	
PPI affinity*	contact order	33	-0.0048	18	$+0.0555$	29	6.2×10^{-4}	yes	18/21	
GO-MF*	composition	32	$+0.0255$	29	$+0.0187$	28	0.001	sign only	18/20	
membrane vs soluble	global shape	33	-0.0156	12	$+0.0011$	28	0.004	sign only	17/21	
membrane vs soluble	secondary structure	33	-0.0124	13	$+0.0058$	28	0.004	sign only	18/21	
membrane vs soluble	size	33	-0.0171	11	$+0.0023$	28	0.004	sign only	18/21	
PPI affinity	size	33	-0.0500	15	$+0.0343$	28	0.004	yes	18/21	
EC classification	global shape	33	$+0.1205$	32	$+0.0149$	27	0.016	sign only	18/21	
fluorescence	contact order	33	-0.0167	4	-0.0022	6	0.016	sign only	2/21	
PPI affinity	global shape	33	-0.0195	16	$+0.0261$	27	0.016	yes	18/21	
solubility	global shape	33	-0.0099	12	$+0.0036$	27	0.016	sign only	19/21	
stability	contact order	33	-0.0556	10	-0.0183	6	0.016	sign only	6/21	
fluorescence*	composition	32	$+0.0087$	20	$+0.0228$	26	0.024	yes	15/20	
stability	composition	32	$+0.0301$	20	$+0.0663$	26	0.024	yes	19/20	yes
fluorescence	global shape	33	-0.0061	21	$+0.0047$	26	0.055		17/21	
GO-MF	contact order	33	$+0.0271$	30	$+0.0037$	26	0.055		16/21	
GO-BP	composition	32	$+0.0120$	24	$+0.0072$	25	0.084		14/20	
solubility	secondary structure	33	-0.0123	9	$+0.0020$	25	0.178		17/21	
membrane vs soluble*	composition	32	-0.0033	13	$+0.0170$	24	0.266		13/20	
fluorescence	secondary structure	33	-0.0216	3	-0.0006	9	0.501		6/21	
PPI affinity	burial and packing	33	-0.0514	14	$+0.0084$	24	0.501		14/21	
EC classification	composition	32	$+0.0888$	32	$+0.0229$	21	1.000		13/20	
EC classification	burial and packing	33	$+0.1171$	32	$+0.0024$	19	1.000		11/21	
EC classification	contact order	33	$+0.1344$	33	$+0.0088$	21	1.000		14/21	
EC classification	secondary structure	33	$+0.1212$	33	$+0.0089$	22	1.000		13/21	
EC classification	size	33	$+0.1275$	33	$+0.0018$	21	1.000		15/21	
fluorescence	size	33	-0.0166	6	$+0.0001$	14	1.000		9/21	
GO-BP	burial and packing	33	$+0.0104$	21	-0.0006	16	1.000		11/21	
GO-BP	contact order	33	$+0.0133$	25	-0.0004	18	1.000		11/21	
GO-BP	global shape	33	$+0.0105$	24	-0.0012	19	1.000		12/21	
GO-BP	secondary structure	33	$+0.0124$	21	$+0.0009$	17	1.000		12/21	
GO-BP	size	33	$+0.0123$	23	-0.0014	13	1.000		9/21	
GO-MF	global shape	33	$+0.0233$	31	$+0.0034$	21	1.000		14/21	
GO-MF	secondary structure	33	$+0.0241$	30	$+0.0014$	21	1.000		14/21	
membrane vs soluble	burial and packing	33	-0.0288	6	-0.0040	12	1.000		8/21	
membrane vs soluble	contact order	33	-0.0188	9	$+0.0002$	16	1.000		10/21	
PPI affinity	composition	32	-0.1050	8	-0.0155	16	1.000		11/20	
remote homology	composition	31	$+0.0002$	14	-0.0026	13	1.000		7/19	
remote homology	burial and packing	32	$+0.0147$	21	-0.0025	16	1.000		10/20	
remote homology	contact order	32	$+0.0156$	19	-0.0017	15	1.000		10/20	
remote homology	global shape	32	$+0.0140$	19	-0.0023	14	1.000		7/20	
remote homology	secondary structure	32	$+0.0188$	19	$+0.0006$	12	1.000		9/20	
remote homology	size	32	$+0.0090$	16	-0.0024	13	1.000		6/19	
solubility	burial and packing	33	-0.0132	12	$+0.0000$	22	1.000		12/21	
solubility	contact order	33	-0.0118	9	$+0.0003$	21	1.000		15/21	
solubility	size	33	-0.0150	9	$+0.0013$	17	1.000		10/21	
stability	burial and packing	33	-0.0311	12	-0.0380	13	1.000		9/21	
stability	global shape	33	-0.0331	13	$+0.0158$	22	1.000		14/21	
stability	secondary structure	33	-0.0313	12	$+0.0058$	15	1.000		9/21	
stability	size	33	$+0.0000$	17	-0.0017	22	1.000		14/21	
subcellular localisation*	composition	32	-0.0389	13	$+0.0080$	21	1.000		11/20	
subcellular localisation	burial and packing	33	-0.0512	12	$+0.0015$	18	1.000		13/21	
subcellular localisation	contact order	33	-0.0474	15	$+0.0002$	15	1.000		9/21	
subcellular localisation	global shape	33	-0.0482	13	$+0.0011$	17	1.000		12/21	
subcellular localisation	secondary structure	33	-0.0517	13	-0.0021	15	1.000		7/20	
subcellular localisation	size	33	-0.0517	14	-0.0026	12	1.000		7/21	

Table 23: Every task-property pair under the repair rule, ordered by the Holm-adjusted p over the 60 pairs. The effect and the repair contrast are both against the variance-matched null; the counts are encoders with a positive value. * named by the standard effect rule. The last two columns are the rule with one vote per architecture family (Appendix D.11): families with a positive family-weighted R_{var}, and whether the pair passes. The pairs on PPI affinity do not hold with chain-disjoint folds (Appendix D.13).

The screened run against the full grid. The 18 pairs of the screened run, recomputed inside the grid with the same code, correlate with it at a median r of 0.93 in their per-encoder repair contrasts, and 17 of the 18 keep their verdict. The one that changes is contact order on PPI affinity, which missed the rule in the screened run (Holm $p = 0.068$) and passes in the grid, with the lowest correlation of the 18 (0.64) on a test split of 326 pairs; it does not hold with chain-disjoint folds (`scripts/grid_reproducibility.py`; the pair-by-pair comparison is released).

D.7 Calibrating the repair rule on labels whose answer is set

Section 3.3 states that the repair test identifies the relevant property in synthetic examples where the answer is known. The positive control of Appendix D.1 checks the erasure effect against the sham; the repair rule uses a different instrument, R_{var} with the floor of 0.02, and needs its own check of power and size: when a property is planted, does the rule name it, and at what strength does it stop? The check also asks about the reach of the variance-matched null, which removes high-variance directions that avoid the property, so that a task using such directions loses score to the null for reasons unrelated to the property.

Labels are built on the stability proteins and twelve encoders, as in the positive control, from the same conditional residual the protocol erases, with the residualisation fitted on the training split and applied to every split (`scripts/matched_null_calibration.py`). Thirteen labels are scored. Nine plant a demand, $y = \sqrt{s}\, f(P) + \sqrt{1-s}\, \varepsilon$, at a *strong*, *weak* or *faint* share $s = 0.5$, 0.05 or 0.01, for each of amino-acid composition, secondary structure content and global shape. Three are *mixed*, one per property, $y = \sqrt{0.05}\, f(P) + \sqrt{0.45}\, g + \sqrt{0.5}\, \varepsilon$. One has nothing planted, $y = \sqrt{0.5}\, g + \sqrt{0.5}\, \varepsilon$. Here $f(P)$ is a fixed linear function of the property's conditional residual, the same for every encoder, and g is a direction of the representation with every property block projected out. A random direction of a representation is dominated by its high-variance components, so g lies in the directions the variance-matched null removes. The label with nothing planted is therefore the condition in which the null costs most, and the mixed labels place a small planted demand inside a task that also uses g. Every map is fitted once per encoder and property, and all thirteen labels are scored against it in one multi-target ridge fit with the penalty chosen per target, which reproduces the regression head of the main analysis to 10^{-6} on a separately fitted column.

	repair contrast vs. matched null				effect of the planted property		
label	planted	positive	largest other	repair rule names	vs. sham	vs. null	effect rule names
strong composition	+0.521	12/12	+0.037	composition	+0.552	+0.501	composition
weak composition	+0.093	11/12	+0.006	composition	+0.163	+0.075	nothing
faint composition	+0.028	12/12	+0.002	composition	+0.050	+0.018	nothing
mixed composition	+0.005	6/12	+0.008	nothing	−0.001	−0.022	nothing
strong sec. structure	+0.446	12/12	+0.088	secondary structure	+0.466	+0.338	secondary structure
weak sec. structure	+0.071	12/12	+0.012	secondary structure	+0.095	+0.038	nothing
faint sec. structure	+0.005	10/12	+0.002	nothing	+0.005	−0.024	nothing
mixed sec. structure	+0.003	6/12	+0.003	nothing	+0.001	−0.030	nothing
strong global shape	+0.488	11/11	+0.121	global shape	+0.555	+0.344	global shape
weak global shape	+0.060	11/11	+0.027	global shape	+0.086	+0.029	global shape
faint global shape	+0.017	11/11	+0.028	nothing	+0.023	−0.005	nothing
mixed global shape	+0.016	9/11	+0.009	nothing	+0.020	−0.029	nothing
none	–	–	+0.008	nothing	–	–	nothing

Table 24: The effect rule and the repair rule on labels whose answer is set, over the encoders whose erasure verified (twelve, eleven for global shape). **Planted** is R_{var} of the planted property and **largest other** the largest R_{var} of any other property under the same label; the last three columns are the planted property's Δ_{sham}, its Δ_{var} and what the effect rule names. Both rules use a two-sided sign test with Holm correction over the eight properties and a floor of 0.02 (`scripts/matched_null_calibration_report.py`).

What the rule does when the answer is known. Table 24 gives every label. On the nine planted labels the repair rule names the planted property at a share of 0.05 or more for all three properties, and at a share of 0.01 for composition, whose repair there (+0.028) clears the floor while secondary structure (+0.005) and global shape (+0.017) do not. Across the thirteen labels, including the one with nothing planted, it names no property that was not planted. The effect rule names the planted property at the strong share, and at the weak share only for global shape. With twelve encoders the sign test is the binding constraint for both rules: ProteinMPNN, which does not read residue identity,

loses nothing on a composition label, and on the weak and faint composition labels the effect rule misses at a Holm-adjusted p of 0.051 while the repair rule passes at 0.044 and 0.004. The rule of Section 3.3 runs over up to 33 encoders, where the sign test is less restrictive.

The reach of the variance-matched null. The null removes high-variance directions that the label with nothing planted uses, so on that label Δ_{var} reaches -0.047 for some property although no property carries any of the label. R_{var} stays within 0.014 of zero for every property on the same label. The mixed labels show where the repair rule is weak. A planted share of 0.05 that the rule names on its own is missed for all three properties when it sits inside a label that mostly uses such directions, and the effect rule misses it as well. The rule is therefore conservative: where a task relies mostly on high-variance directions that avoid the property, it can miss a demand. Subcellular localisation behaves this way in Section 3.3.

The scale of the floor. The floor was fixed before any repair contrast existed. The descriptor with its rows permuted across proteins, the same width and marginal with no protein information, gives a panel-mean R_{var} of at most 0.0066 in absolute value on every task but EC classification, so on those nine tasks the floor is three times what a descriptor unrelated to the protein produces. On EC classification and stability the permuted descriptor is noisier (resampled panel means up to 0.029 and 0.021), so there the sign test carries the decision; no pair on EC classification is named, and on the seven named pairs the permuted descriptor's own contrast lies between -0.0021 and $+0.0001$ (`scripts/repair_permuted_null.py -tag full`). On the calibration labels the floor sits between a planted share of 0.01 and 0.05.

D.8 Does the named property repair its own erasure? A selective add-back

Section 3.3 attributes an erasure's loss to a property by adding the property back. Is that repair selective, so that the named descriptor repairs its own erasure while a permuted copy of it and the other property blocks do not? Equation 1 of Section 2.4 defines the repair contrast as the gain an independently computed descriptor q brings to the erased representation minus its mean gain over the members of a null family, here the sham family (R_{sham}) and the variance-matched family (R_{var}). The descriptor is computed from the protein and never from the label. Three descriptors are compared at every arm: the named property; the same columns with the proteins permuted, which match the width and the marginal and carry no relation to the protein; and every other independent structural block. The test runs on the three tasks the effect rule names for amino-acid composition, over six encoders (ESM-2 650M, ESM-1b, ProtT5-XL, SaProt, ESM-IF1 and ProteinMPNN), with four sham permutations per cell and every map, standardiser and head fitted on the training split (`scripts/addback_rescue.py`).

Table 25: Selective repair. **Intact / erased / sham** are the mean scores over encoders with no descriptor added. **Mean** R is R_{sham} of Equation 1 averaged over encoders, with the spread over encoders, and $R > 0$ counts the encoders on which the descriptor repairs the real erasure more than it repairs the shams. The head is the linear one; Appendix D.12 gives the two-layer head.

task	descriptor added back	encoders	intact / erased / sham	mean R	$R > 0$
membrane vs soluble	amino-acid composition (named)	6	0.868 / 0.824 / 0.860	$+0.0362 \pm 0.004$	6 / 6
	the same, proteins permuted	6		$+0.0000 \pm 0.001$	2 / 6
	the other structural blocks	6		-0.0018 ± 0.002	1 / 6
solubility	amino-acid composition (named)	6	0.660 / 0.636 / 0.656	$+0.0180 \pm 0.006$	6 / 6
	the same, proteins permuted	6		$+0.0002 \pm 0.000$	4 / 6
	the other structural blocks	6		$+0.0028 \pm 0.004$	6 / 6
subcellular localisation	amino-acid composition (named)	6	0.606 / 0.583 / 0.599	$+0.0146 \pm 0.002$	6 / 6
	the same, proteins permuted	6		-0.0003 ± 0.002	2 / 6
	the other structural blocks	6		$+0.0003 \pm 0.003$	3 / 6

The named descriptor repairs its own erasure and the controls do not (Table 25): $R_{\mathrm{sham}} = +0.036$ on membrane and soluble, $+0.018$ on solubility and $+0.015$ on subcellular localisation, positive on 6 of 6 encoders on each task, while the permuted copy stays within 0.001 of zero and the other structural blocks within 0.003. Against the variance-matched family, which removes as much of the representation as the real erasure, R_{var} reads $+0.019$, $+0.013$ and $+0.007$ (positive on 5, 6 and 5 of

6 encoders) and the controls stay near zero (the permuted copy from -0.0014 to $+0.0002$, the other blocks from -0.0145 to $+0.0039$), while the erasure effect on the same cells changes sign (Δ_{sham} $+0.036$, $+0.019$ and $+0.014$ with every encoder positive; Δ_{var} -0.018, -0.011 and -0.037 with one encoder of six positive). A repair is a difference of differences, so it cancels the capacity loss the two arms share, which is why the repair rule rests on it. The test uses the same proteins as the rest of the paper; Appendix E.4 repeats it on proteins never scored (hypothesis D2). A positive R shows that the erased representation lacks something the composition descriptor supplies, and it leaves open whether other information was removed together with composition.

D.9 A FIXED HEAD: DOES THE FITTED PREDICTOR RELY ON THE ERASED DIRECTIONS?

The erasure analysis of Section 3.3 retrains the head after every intervention. Three quantities come from the same interventions and answer different questions. The *retrained effect* Δ_{var} asks how much predictive information stays accessible after the property's subspace is removed and a new head adapts; the *repair contrast* R_{var} asks whether adding the property back restores its own erasure more than a variance-matched removal; the *fixed-head effect* $\Delta^{\text{fix}}_{\text{var}}$ asks whether the head fitted on the intact representation relies more on the erased directions than on matched ones. For the last, the head is trained once on the intact training rows and scored on test inputs projected in the original coordinates, $x' = \mu + (x - \mu)P^\top$, with the erasure, sham and variance-matched null projections of the released grid, and $\Delta^{\text{fix}}_{\text{var}}$ is the mean over the four nulls of $S(h(x'_{\text{null}})) - S(h(x'_{\text{erased}}))$. The decision rule is the effect rule applied to $\Delta^{\text{fix}}_{\text{var}}$, with Holm correction, and was fixed before any fixed-head score was computed (`erasure/ab_fixed_head/PRESPEC.md`; `scripts/ab_fixed_head.py`). It covers the eleven pairs that the effect rule or the repair rule names, 357 encoder cells. The projection equals the released erased arm up to an isometry, and the frozen head reproduces the released intact score to within 0.0008 outside PPI affinity and 0.022 on it.

The fixed-head rule names three pairs (Table 26): burial and packing on fluorescence ($+0.056$, 33 of 33 encoders), composition on fluorescence ($+0.031$, 26 of 32) and composition on solubility ($+0.021$, 30 of 32, 0.0005 above the floor); all three stay named under Holm correction over the 60 pairs. Composition on stability, the pair that survives the follow-ups of Section 3.3, has a positive mean ($+0.047$) but is positive on 21 of 32 encoders (Holm $p = 0.22$): adding composition back repairs its erasure selectively, while the fitted heads do not rely on the composition directions consistently across encoders. On PPI affinity the fixed-head effect is negative on all four pairs the repair rule names, significantly so for secondary structure (6 of 33 positive), and for secondary structure and global shape the frozen head scores higher on erased inputs than on intact ones (0.434 and 0.449 against 0.411). The three quantities agree on the two fluorescence pairs and disagree on most others, which is what separating them is for: a property can be recoverable after removal, selectively repaired, or relied on by a fitted head without the other two holding.

		fixed head				retrained head (released)		
task	property	n	$\Delta^{\text{fix}}_{\text{var}}$	positive	Holm p	Δ_{var}	R_{var}	rules
fluorescence	burial and packing	33	$+0.0555$	33	2.6×10^{-9}	$+0.0182$	$+0.0218$	E+R+F
PPI affinity	secondary structure	33	-0.0435	6	0.003	$+0.0188$	$+0.0721$	E+R
solubility	composition	32	$+0.0205$	30	2.5×10^{-6}	$+0.0069$	$+0.0193$	E+F
PPI affinity	contact order	33	-0.0080	14	0.487	-0.0048	$+0.0555$	E+R
GO-MF	composition	32	$+0.0098$	23	0.100	$+0.0255$	$+0.0187$	E
PPI affinity	size	33	-0.0106	8	0.027	-0.0500	$+0.0343$	R
PPI affinity	global shape	33	-0.0545	7	0.009	-0.0195	$+0.0261$	R
fluorescence	composition	32	$+0.0310$	26	0.004	$+0.0087$	$+0.0228$	E+R+F
stability	composition	32	$+0.0472$	21	0.220	$+0.0301$	$+0.0663$	R
membrane vs soluble	composition	32	$+0.0197$	23	0.100	-0.0033	$+0.0170$	E
subcellular localisation	composition	32	$+0.0178$	22	0.150	-0.0389	$+0.0080$	E

Table 26: Fixed-head and retrained-head erasure effects and the repair contrast on the eleven named pairs. Fixed head: the head fitted on the intact representation, scored on test inputs projected in the original coordinates; $\Delta^{\text{fix}}_{\text{var}}$ is the mean over four variance-matched nulls of $S(\text{null}) - S(\text{erased})$, with a two-sided sign test over encoders and Holm correction over the eleven pairs. Retrained columns are the released Δ_{var} and R_{var}. Rules: E effect rule, R repair rule, F fixed-head rule.

D.10 COMPETING EXPLANATIONS ON FLUORESCENCE AND STABILITY

Outside PPI affinity the repair rule names three pairs (Section 3.3): burial and packing and amino-acid composition on fluorescence, and composition on stability. Both tasks score variants of a few parent proteins. This subsection asks whether each descriptor repairs the erasure through the named property or by standing in for simpler features of the mutations.

Every fluorescence variant is a copy of one parent protein with a few substitutions. Before computing the result we fixed a competing set of mutation features (`protcompass/qd/gfp_competing/PRESPEC.md`): the number of edits to the parent with its square and bins, eleven substitution-class counts (hydrophobic gained or lost, to or from proline and glycine, charge-altering, net charge change, aromatic gained or lost, gaps), the mean pLDDT of the predicted structure and, for burial and packing, the change in amino-acid counts. The change in counts is left out for composition, because for variants of the parent's length it determines the composition descriptor. The competing features enter every arm, so the repair contrast measures what the descriptor adds beyond them (`protcompass/scripts/gfp_competing.py`); the arms with and without them come from one run.

The competing features alone reach a test Spearman correlation of 0.49. With them in every arm, the repair contrast falls from $+0.025$ (in this run; the grid of Table 23 gives $+0.023$) to $+0.004$ for composition (25 of 32 encoders positive, Holm $p = 0.004$; 18 of 20 families) and from $+0.022$ to $+0.0003$ for burial and packing (24 of 33 encoders, Holm $p = 0.014$; 16 of 21 families). Both stay positive in sign and both fall below the 0.02 floor. The predicted structures of the variants all come from one predictor, so structural descriptors of them are themselves partly functions of the substitutions. On fluorescence the two named properties therefore mostly carry information that simple statistics of the mutations already contain.

Stability poses the same question. Its test split is 2,998 single mutants of 17 parent proteins, 13 of which never appear as a training parent, so composition could act as a label for the parent. The pre-specified competitors (`protcompass/qd/stability_competing/PRESPEC.md`) are the edits to the parent with square and bins, the eleven substitution-class counts, mean pLDDT, and the topology, design round and variant kind of each protein. Two candidates are left out: the change in amino-acid counts, which for single mutants is the descriptor itself, and a parent indicator, which cannot transfer to parents unseen in training. A second score, the Spearman correlation within each parent averaged over the 17, removes parent identity by construction (`scripts/stability_competing.py`). The run without competitors reproduces the grid cells exactly. The competitors alone reach a Spearman correlation of 0.20. With the competitors in every arm the composition repair rises from $+0.066$ to $+0.100$ (30 of 32 encoders, 20 of 20 families), the permuted descriptor stays at $+0.002$, and within parents it is $+0.030$, positive on all 32 encoders. Composition on stability therefore passes the repair rule with these competitors in every arm, and it is the one demand in this paper that survives the controls and competitors we ran. The pair was named by the repair rule after hypothesis D3 had predicted that nothing would be named on stability (Appendix E.4), so its held-out support is not a frozen prediction.

Which part of composition carries the stability repair? For a single mutant, the composition descriptor is the parent's background composition plus the substitution vector $(e_{\text{mutant}} - e_{\text{wild type}})/L$: across parents it identifies the parent, and within a parent it encodes the substitution, so the repair above could come from either. We split the descriptor exactly into the two parts, the parent background (the variant's composition with every identified substitution reverted) and the substitution vector, which sum to the released descriptor. Each part is added back alone, and each is also added back with the other part present in every arm. The released competitors (C0) are used, and a larger set (C1) adds the wild-type and mutant residues of every substitution, its relative position, and the relative solvent accessibility and DSSP class of the substituted site on the ESMFold structure. The erasure, nulls, heads, encoders and decision rule are those of the released run, which the new run reproduces exactly. The analysis was specified before scoring, with an addendum for the conditional rows and the parent-level intervals written before those rows were computed (`erasure/ab_stability_decomp/PRESPEC.md`, `PRESPEC_addendum.md`).

The repair is carried by the parent background (Table 27). The substitution vector repairs nothing, whether it is added alone or with the parent background in every arm, while the parent background keeps a repair of $+0.085$ on 29 of 32 encoders and 19 of 20 families with the exact wild-type and

Table 27: The stability composition repair split into its two parts. R_{var} is the mean over 32 encoders of the pooled-Spearman repair contrast against the variance-matched null; the counts are encoders (of 32) and architecture families (of 20) with a positive contrast. The within-parent column is the Spearman contrast inside each of the 17 test parents, averaged over parents, with a 95% bootstrap interval over parents. C0: released competitors; C1: C0 plus the wild-type and mutant residues, the position and the site environment of each substitution.

descriptor added back	competitors	R_{var}	encoders, families	within parent [95% CI]
full composition (released)	C0	$+0.100$	30, 20	$+0.030$ $[+0.017, +0.044]$
	C1	$+0.085$	29, 19	$+0.013$ $[+0.005, +0.021]$
parent background	C0	$+0.099$	31, 20	$+0.034$ $[+0.025, +0.046]$
	C1	$+0.085$	29, 19	$+0.014$ $[+0.007, +0.022]$
substitution vector	C0	$+0.003$	16, 9	$+0.001$ $[-0.005, +0.005]$
	C1	$+0.002$	16, 10	-0.002 $[-0.006, +0.002]$
parent background, substitution in every arm	C1	$+0.085$	31, 19	$+0.015$ $[+0.007, +0.023]$
substitution, parent background in every arm	C1	$+0.002$	12, 9	-0.001 $[-0.005, +0.002]$

mutant residues and the site environment in every arm. Without any representation, the competitors reach a Spearman correlation of 0.204 (C0) and 0.259 (C1); adding the parent background raises these to 0.333 and 0.401, and adding the substitution vector does not raise them (0.183 and 0.254). Thirteen of the 17 test parents never appear as a training parent, so the background composition acts as a transferable description of the design a variant comes from. The composition demand on stability is therefore a demand for information about the parent design's sequence background, and the substitution made in the variant does not contribute to it.

D.11 The repair rule with task-calibrated floors, well-matched cells and one vote per family

The repair rule uses a magnitude floor of 0.02 on every task's score scale, although the ten tasks use four metrics, and it includes cells whose variance-matched null matches the removed variance only loosely. This subsection asks whether the named pairs depend on either choice, and whether they depend on counting a repeated architecture several times; Section 3.5 reports that composition on stability remains named under these variants. We recomputed the rule on the full grid with one element changed at a time, which gives four variants besides the rule itself (`scripts/naming_sensitivity.py`, Table 28). Two task-calibrated floors replace 0.02: the task's minimum detectable effect from the sham spread ($\alpha = 0.05$, power 0.8, the largest over the properties erased on that task), and 5% of the task's reachable range (Appendix D.5). A third variant keeps 0.02 and uses only the cells whose variance-matched null removes the erased share of variance to within 1%, which holds for 1,245 of the 1,964 cells of the 60-pair grid (63.4%; Appendix D.4). The fourth is the family rule described below.

Composition on stability is named under every variant. Composition on fluorescence is named under four of the five and burial and packing under three, and the PPI pairs under three to five; all of them fail their competing-explanation tests (Appendices D.10 and D.13) whichever variant names them. The calibrated floors are lower than 0.02 on GO-MF and solubility and add composition on both tasks ($+0.019$ each), and under one floor each, global shape on EC ($+0.015$) and size on GO-MF ($+0.008$). These repairs are below 0.02 on the score scale and no competing explanation has been tested on them, so we report them as candidates.

One vote per architecture family. The sign test runs over encoders, and the 33 encoders are not independent: ten of the 23 coordinate encoders are randomly initialised, five of them random versions of pretrained networks also in the panel, ESM-2 appears at five scales and PottsMPNN at four. The family rule first averages the encoders inside an architecture family, a random and a pretrained run of one network being one family, and runs the sign test over the 21 families, with a percentile bootstrap interval over families and the verdict repeated with each family left out (`scripts/family_weighted_repair.py`); its verdicts are the last two columns of Table 23. It names three of the 60 pairs: burial and packing on fluorescence (21 of 21 families, $+0.0215$), composition on stability (19 of 20, $+0.0732$) and secondary structure content on PPI affinity (20 of 21), the last of which does not hold with chain-disjoint folds (Appendix D.13). Stability composition stays

named with every one of the 21 families left out, and fluorescence burial with 20 of them. Composition on fluorescence, which the rule over encoders names, is positive in 15 of 20 families at $+0.0195$, just under the floor. Composition on solubility (20 of 20 families) and GO-MF (18 of 20) keeps its sign across families, and composition on membrane vs soluble and GO-BP does not (13 and 14 of 20), so the consistency seen over encoders there comes from repeated architectures. Under the standard effect rule the same reweighting keeps seven of its eight named pairs and loses GO-MF composition on the magnitude floor ($+0.0202 \rightarrow +0.0187$; `scripts/family_weighted_demand.py`). Other dependence remains, since encoders share pretraining corpora and proteins and the families are our grouping, so the statements of Section 3.3 describe this panel.

Table 28: Every task-property pair that at least one variant of the repair rule names, with the mean repair contrast in that variant. The matched variant averages over the well-matched cells and the family variant over architecture families, so their means differ from the released one (`erasure/naming_sensitivity.csv`).

task	property	released (0.02)	task MDE	5% of reach	matched cells	families
stability	composition	**yes** ($+0.066$)	**yes** ($+0.066$)	**yes** ($+0.066$)	**yes** ($+0.066$)	**yes** ($+0.073$)
EC	global shape	no ($+0.015$)	**yes** ($+0.015$)	no ($+0.015$)	no ($+0.014$)	no ($+0.015$)
fluorescence	composition	**yes** ($+0.023$)	**yes** ($+0.023$)	**yes** ($+0.023$)	**yes** ($+0.027$)	no ($+0.020$)
fluorescence	burial and packing	**yes** ($+0.022$)	**yes** ($+0.022$)	no ($+0.022$)	no ($+0.018$)	**yes** ($+0.022$)
GO-MF	composition	no ($+0.019$)	**yes** ($+0.019$)	**yes** ($+0.019$)	no ($+0.019$)	no ($+0.018$)
GO-MF	size	no ($+0.008$)	no ($+0.008$)	**yes** ($+0.008$)	no ($+0.006$)	no ($+0.009$)
PPI affinity	contact order	**yes** ($+0.056$)	**yes** ($+0.056$)	**yes** ($+0.056$)	**yes** ($+0.059$)	no ($+0.050$)
PPI affinity	global shape	**yes** ($+0.026$)	no ($+0.026$)	**yes** ($+0.026$)	**yes** ($+0.026$)	no ($+0.021$)
PPI affinity	secondary structure	**yes** ($+0.072$)	**yes** ($+0.072$)	**yes** ($+0.072$)	**yes** ($+0.072$)	**yes** ($+0.075$)
PPI affinity	size	**yes** ($+0.034$)	no ($+0.034$)	**yes** ($+0.034$)	**yes** ($+0.031$)	no ($+0.035$)
solubility	composition	no ($+0.019$)	**yes** ($+0.019$)	**yes** ($+0.019$)	no ($+0.019$)	no ($+0.019$)

D.12 Composition on stability with a nonlinear head and a nonlinear erasure

Composition on stability is the one pair that survives the controls and competitors we ran, including the exact wild-type and mutant residues, and its signal is carried by the parent background composition (Appendix D.10); the demand named on it is a linear-head demand. This subsection asks whether it holds when the head, or the erasure and the head together, are nonlinear in the representation; Sections 3.3 and 3.5 report the two results. We ran two nonlinear versions of the repair test on it (`scripts/stability_nonlinear_summary.py`, Table 29).

A nonlinear head. LEACE leaves most of the property readable by a two-layer network (Appendix D.2), so an MLP head can use what a linear erasure leaves behind. With the downstream MLP of the benchmark (two hidden layers of width 512) as the head in every arm, on twelve encoders (ESM-2 650M and 150M, ESM-1b, Ankh, ProtT5, ProstT5, ESM3, SaProt, ESM-IF1, ProteinMPNN, GVP and GearNet (pretrained)), the repair contrast has a mean of $+0.037$ and is positive on 7 of 12 encoders, against $+0.073$ and 10 of 12 with the linear head on the same encoders. The contrast loses its consistency in sign, as it does for the other named pairs: recomputed with the two-layer head on the same twelve encoders for composition on six tasks, burial and packing on fluorescence and secondary structure content on PPI affinity (`scripts/mlp_repair_summary.py`), no pair passes the repair rule. Composition on fluorescence keeps its direction and about 0.6 of its size ($+0.019$ on 10 of 12 encoders, against $+0.032$ on 11 of 12 with the linear head), and burial and packing about a tenth ($+0.003$ on 8 of 12, against $+0.026$ on 12 of 12). The selective add-back of Appendix D.8 behaves the same way: with a two-layer head its R_{sham} falls inside the noise of the controls ($+0.001 \pm 0.007$, $+0.010 \pm 0.030$ and $+0.012 \pm 0.010$). Appendix D.2 predicts this: the erasure leaves a median 0.86 of the property recoverable by a two-layer readout, so a nonlinear head reconstructs much of what the descriptor would supply.

A nonlinear erasure. Every representation is mapped to 2,048 random Fourier features of a Gaussian kernel on its first 256 principal components (bandwidth the median pairwise distance on the training split), and LEACE, the sham, the variance-matched null, the descriptor and the linear head are all applied in that feature space (`scripts/addback_kernel.py`). The head is then a kernel machine on the original representation, and the erasure removes what such a machine can read linearly from the features. On six encoders, a linear readout of the conditional composition from the features falls

from a median R^2 of 0.84 to -0.09 after this erasure, while a two-layer MLP on the erased features still reaches a median of 0.63 (`erasure/named_pairs_kernel/recoverability.csv`), so a more flexible head can still read part of the property. With this erasure the repair contrast is $+0.085$ and positive on 32 of 33 encoders; Foldseek 3Di enters only here, because its 20-dimensional representation leaves too few directions after the linear erasure. The permuted descriptor gives $+0.0009$. Composition on stability therefore holds when the erasure and the head are both nonlinear in the representation and matched to each other, within the class the kernel defines.

Table 29: Composition on stability: the repair contrast R_{var} with a two-layer head and after a kernel-space erasure, each against the linear result on the same encoders (`erasure/stability_nonlinear.csv`).

variant	encoders	R_{var}	positive	sign test p	permuted descriptor
two-layer head	12	$+0.037$	7	0.774	-0.0012
linear head, same encoders	12	$+0.073$	10	0.039	$+0.0002$
kernel-space erasure	33	$+0.085$	32	< 0.001	$+0.0009$
linear erasure, same encoders	32	$+0.066$	26	< 0.001	-0.0021

D.13 PPI AFFINITY WITH CHAIN-DISJOINT FOLDS

Four pairs on PPI affinity pass the repair rule over the full grid (Appendix D.6): secondary structure content, contact order, size and global shape. Section 3.3 attributes them to chains shared between training and test, and this subsection gives the test. The released PPI affinity split shares chains between training and test: every one of its 326 test pairs contains a chain that also appears in a training pair, and one chain appears in 283 pairs. A head can therefore score on a test pair by recognising a chain it was trained on, and a descriptor of the two chains, added back after an erasure, can help it do so whether or not the task uses the property the descriptor measures.

We re-measure all four on the 2,872 pairs with folds that share no chain (`scripts/ppi_repair_cv.py`). Treating chains as nodes and pairs as edges gives 72 connected components, and whole components are assigned to five folds, three times with different assignments. Inside each fold every fitted step, from the PCA of the released pair loader to the conditional target, the LEACE map, the four shams, the four variance-matched nulls and the released regression head, is fitted on the training part alone. Each encoder's contrast is its mean over the 15 folds, and the repair rule is applied to those means, with Holm correction over the four properties. The interval is a percentile bootstrap that resamples encoders and, inside each encoder, folds. As a control for the chain constraint itself, the same five folds and three repeats were drawn with pairs assigned at random, so that test pairs share chains with training pairs as in the released split, with everything else unchanged (`-random-folds`).

Table 30: The repair contrast against the variance-matched null on PPI affinity with chain-disjoint folds (five folds, three assignments), for the four pairs the full grid names. **Families** counts the architecture families with a positive family mean.

property	encoders	R_{var} [95% CI]	Holm p	families	R_{var} by assignment	passes
contact order	11/33	-0.0211 [-0.0405, -0.0029]	0.24	6/21	+0.0058 / -0.0419 / -0.0273	no
global shape	12/33	-0.0107 [-0.0298, $+0.0075$]	0.326	8/21	+0.0131 / -0.0289 / -0.0163	no
secondary structure content	19/33	$+0.0114$ [-0.0062, $+0.0292$]	0.487	15/21	+0.0272 / -0.0165 / +0.0226	no
size	10/33	-0.0128 [-0.0269, $+0.0016$]	0.14	7/21	-0.0218 / -0.0090 / -0.0075	no

None of the four is named with chain-disjoint folds (Table 30). Secondary structure content repairs by $+0.011$ [-0.006, $+0.029$], positive on 19 of 33 encoders, and its mean changes sign between the three fold assignments; contact order (-0.021), size (-0.013) and global shape (-0.011) are negative. With random folds of the same design, secondary structure content repairs by $+0.022$ on 32 of 33 encoders and passes the rule (`erasure/ppi_repair_cv_random_summary.csv`). Random folds have the same smaller training sets as chain-disjoint folds, so the loss with chain-disjoint folds comes from the chain constraint. The four PPI pairs therefore describe the released split, in which test pairs share chains with training pairs. Outside PPI affinity the repair rule names three pairs, of which composition on stability survives the competing explanations of Appendix D.10.

E CONFIRMATION ON DATA AND SPLITS THE DEVELOPMENT DID NOT USE

The floors, descriptors and naming rules were developed on the superfamily split, so their results could be fitted to it. This section asks whether the results of Sections 3.1 and 3.3 hold where the development did not look. Each test had its hypotheses fixed before any score on its data existed, and each is reported as its frozen rules decide it (Table 31).

Table 31: The four confirmation tests.

test	what is new	hypotheses	outcome	appendix
topology-disjoint split	folds absent from training: 1,924 test domains in 122 topologies	T1–T4	all four hold	E.1
CATH 4.4 cohorts A and B	domains absent from CATH 4.3: 1,990 and 1,316	H1–H4; H1b, H2b, H5 on B	H2, H2b and H3 hold; H1 and H4, whose thresholds the released split itself does not meet, do not	E.2
pre-registered labels	three residue labels frozen before any encoder was scored on them	H1–H4	H1, H3 and H4 hold; H2 does not	E.3
downstream held-out	13,632 downstream proteins never extracted or scored	D1–D4	D1 and D2 hold; D3 and D4 undecided	E.4

E.1 A FROZEN TEST ON FOLDS ABSENT FROM TRAINING

The floors and the determinacy reading of Section 3.1 were developed on the superfamily split. The CATH 4.3 topology split holds 1,924 test domains in 122 topologies, none of which occurs in its training split (15,378 domains). Before any source was scored on it, we froze the panel, the floors, the protocol of Appendix B.2 and four hypotheses, with thresholds checked against the superfamily split so that a faithful replication could meet them. Fold architecture was the only label scored on this split before the freeze, and no floor had been scored on it.

Table 32: The four hypotheses frozen before the topology-disjoint split was scored, as the frozen rules decide them (`symmetric/hypotheses_topology.json`). The frozen wording calls the input-matched floor admissible.

	hypothesis, as frozen	result	
T1	Of the coordinate-encoder cells on the eleven determined labels, at least 85% are below the admissible floor (95% interval below zero).	234 of 245 (96%)	holds
T2	No sequence-only cell on the eleven determined labels is below its admissible floor (95% interval below zero).	0 of 110	holds
T3	On each of RSA, RSA bins and B-factor, at least one encoder is above its admissible floor (95% interval above zero).	RSA: 20; RSA bins: 20; B-factor: 13	holds
T4	Over coordinate-encoder cells, the mean gap to the admissible floor is higher on the three undetermined labels than on the eleven determined ones, with a 95% bootstrap interval over architecture families (10,000 draws) above zero.	$+0.178$ $[+0.146, +0.217]$	holds

All four hypotheses hold (Table 32). On the eleven determined labels 234 of the 245 coordinate-encoder cells are below the input-matched floor and 11 are above, the same eleven encoder-label pairs as on the superfamily split, and none of the 110 sequence-only cells is below (108 above, two unresolved); the topology rows of Table 6 give the counts under every floor rule, including the label-agnostic family. The gap is higher on undetermined than on determined labels by $+0.178$ $[+0.146, +0.217]$ over 15 architecture families, against $+0.186$ on the superfamily split, with ESM3 above every floor on RSA by $+0.062$ $[+0.061, +0.063]$ and ESM-IF1 on B-factor by $+0.071$ $[+0.064, +0.078]$; the fold-architecture margins on this split are those of Appendix B.8. Selectivity on DSSP-3 again orders the encoders against the floors, and the four frame-projected equivariant encoders stay below the floor on all 44 of their determined cells. Two of the four predictions are

largely set by the design. T1 needs 209 cells below the floor, and the seven labels where a floor is near its ceiling or contains the label, together with the randomly initialised networks on the other four, give 196. T2 asks floors built from the sequence, which cannot compute these labels, to stay below a sequence encoder. T3 and T4 are the predictions the new folds could more easily have failed.

E.2 The probing headline on CATH 4.4 domains absent from CATH 4.3

Do the probing results hold on proteins that no descriptor in this paper was designed against? Two cohorts of domains of the CATH v4.4 non-redundant S40 set that are absent from the whole of v4.3 were scored once, each with its panel, protocol and hypotheses frozen before scoring. Cohort A is 1,990 such domains, none of which appears in the released dataset, its splits or its rejection log. Cohort B is every other domain of that set, 1,324 domains in 20 architectures (1,316 after label construction), fixed before any structure in it was extracted. Both use the panel ESM2-650M, ProtT5, Ankh, ESM3, ESM-IF1 and ProteinMPNN and the released seed-0 fitting draw of 3,000 training and 500 validation domains, verified index-identical to the draw the released scorer makes. Cohort B repeats cohort A's four hypotheses unchanged and adds three: H1b and H2b hold each encoder only to input-matched floors chosen on validation (Appendix B.1), and H5 asks for H1 to H4 on the domains whose topology the released training and validation splits do not contain.

Table 33: The probing hypotheses of the two CATH 4.4 cohorts, as the rules frozen before scoring decide them. H1b, H2b and H5 were added for cohort B.

	prediction, fixed before the cohort was scored	result on cohort A	verdict	result on cohort B	verdict
H1	On the nine informative determined probes, every coordinate-reading encoder in the panel scores below the floor built from the atoms it receives.	coordinate encoders below their floor on the nine informative determined probes: esm3 6/9, esm_if1 7/9, proteinmpnn 8/9	not confirmed	per encoder below of 9: esm3 6, esm_if1 7, proteinmpnn 8	not confirmed
H2	On the nine informative determined probes, no sequence-only encoder in the panel falls below the sequence floor.	sequence-only cells below the sequence floor: 0 of 27	confirmed	0 of 27 sequence-only cells below	confirmed
H3	On RSA, RSA bins and B-factor, some encoder in the panel beats every floor.	best encoder minus best floor: B-factor +0.203, RSA +0.255, RSA bins +0.168	confirmed	best gap: bfactor +0.197, rsa +0.263, rsa_bins +0.174	confirmed
H4	On fold architecture, ESM-IF1, ESM3 and ProstT5 clear ref_pairhist and ESM2-650M does not. ProstT5 is outside this panel, so the prediction is tested for ESM-IF1 and ESM3.	against ref_pairhist on fold architecture: esm_if1 +0.302, esm3 +0.381, esm2_650m +0.339	not confirmed	vs ref_pairhist: esm_if1 +0.291, esm3 +0.384, esm2_650m +0.309	not confirmed
H1b	H1 with each encoder held only to floors built from inputs it is measured to read (the residue-identity audit and the atom-level audit), the floor descriptor and readout family chosen on the released validation split.	–		per encoder below of 9: esm3 6, esm_if1 6, proteinmpnn 8	not confirmed
H2b	H2 with the sequence floor chosen on the released validation split.	–		0 of 27 sequence-only cells below	confirmed
H5	H1 to H4 hold on the topology-novel part of cohort B: domains whose CATH topology is absent from the released training and validation splits.	–		47 domains; H1 False (esm3:6 esm_if1:7 proteinmpnn:8), H2 True (0 of 27), H3 True (bfactor:+0.158 rsa:+0.269 rsa_bins:+0.148), H4 False	not confirmed

H2 and H3 confirm on both cohorts, and so does H2b; H1 and H4 do not (Table 33). H2 predicts that no sequence-only encoder falls below the sequence floor on the nine informative determined probes, and H3 that some encoder beats every floor on the undetermined labels. H1 and H4 set thresholds that the released split itself does not meet. H1 requires every coordinate encoder to be below its floor

on at least seven of the nine probes, and ESM3, ESM-IF1 and ProteinMPNN are below on 6, 7 and 8 on the released split and on both cohorts. H4 requires ESM2-650M to stay below `ref_pairhist` on fold architecture, which it clears on the released split by $+0.181$ and on the cohorts by $+0.339$ and $+0.309$. H1b fails with H1 (under the input-matched floors ESM-IF1 moves from 7 to 6), and H5 fails with H1 and H4, while on its 47 new-topology domains H2 and H3 hold. The frozen test on the topology-disjoint split (Appendix E.1) checked its thresholds against the development split before freezing, and all four of its hypotheses hold.

The cohorts reproduce the released split encoder by encoder. Each coordinate encoder passes the floor on the same probes as on the released split (core/surface, DSSP-3 and DSSP-8 for ESM3, core/surface and DSSP-3 for ESM-IF1, core/surface for ProteinMPNN), and on cohort A the scores of every floor and every encoder differ from the released run by 0.001 to 0.02, except on fold architecture, where the descriptor floor falls from 0.316 to 0.220 while the encoders rise.

New proteins of known folds, and new folds. Every topology among the domains added between v4.3 and v4.4 already exists in v4.3: 1,802 of the 1,990 domains of cohort A share a topology with the fitting set, 1,428 share a homologous superfamily, and three carry a sequence identical to a fitting domain. The cohorts therefore test new proteins of known folds, and the topology-disjoint split tests new folds. The fits were also scored on the parts of cohort A whose topology or homologous superfamily is absent from the fitting draw (188 and 562 domains) or from the whole released training and validation split (51 and 218 domains) (`scripts/confirmation_novelty.py`). Every hypothesis is decided as on the whole cohort: in every subset the coordinate encoders fall below the floor on 21 of 27 cells and the sequence-only encoders on 0 of 27, and the fold-architecture margin over `ref_pairhist` shrinks as the folds become newer, from $+0.302$ to $+0.072$ for ESM-IF1 and from $+0.381$ to $+0.072$ for ESM3 on the 51 domains whose topology the released split does not contain, and stays positive. The cohorts do not control for pretraining exposure.

E.3 A PRE-REGISTERED TEST OF DETERMINACY ON NEW LABELS

Section 3.1 reads the floor comparison through determinacy: where the encoder's input fixes the label, the encoder sits at or below a floor built from that input, and where it does not, some encoder passes every floor. That reading was formed on the 21 released probes. Does it predict the result on labels it was not formed on? We froze an empirical prediction before any encoder or floor was scored on new labels (`prereg/freeze.json`, with the sha256 of every script and label file it depends on). The prediction is stated for the frozen floor family and protocol: labels determined by an encoder's audited input are predicted to give predominantly non-positive gaps G_{ref} (encoder minus floor), and labels not determined by it predominantly positive ones.

Labels and determinacy. Three primary labels were computed from the domain files: the number of backbone hydrogen bonds a residue takes part in (DSSP energy; classes 0, 1, 2 or more), the χ_1 rotamer class and the χ_2 rotamer class (Arg, Lys, Gln, Glu, Met, Ile, Leu and Trp), with the Cα virtual dihedral as a positive control and the side-chain contact count as a descriptive label (`scripts/prereg_labels.py`, `prereg_chi2.py`); candidates that repeat released probes were dropped before freezing. Determinacy has three states (`scripts/prereg_determinacy.py`): a cell is *determined* if the label is reconstructed exactly from the atoms the encoder is measured to read, *not determined* if an executable construction changes the label while every atom it reads stays fixed (a peptide-plane rotation about Cα–Cα for the H-bond count, a side-chain rotation about Cα–Cβ or Cβ–Cγ for the rotamers), and *not established* otherwise. Reconstruction agreed on every residue of 200 domains where it was claimed. Placing O ideally from N, Cα and C reproduces the H-bond count on a mean of only 0.975 of residues, so the encoders that read N, Cα and C, and those that read N, Cα and Cβ, are not established on that label and stay out of the test. Both GCPNets and FoldVision read atoms beyond Cβ (`extract/sidechain_audit.py`); FoldVision has no residue-level embedding and is not probed.

Floors, protocol and decision rules. Each encoder is held to every floor built only from inputs it reads, including two generic descriptors added before freezing, a non-local atom-neighbour descriptor on the encoder's own atom set and a residue-level all-atom descriptor for the side-chain readers; for encoders and floors alike the readout family and the floor are chosen on validation and the test score is read once (`scripts/prereg_floors.py`, `prereg_run.py`, `prereg_decide.py`). Units are

architecture families, with G_{ref} averaged over a family's members of the same status. H1: at least 80% of determined units have $G_{\mathrm{ref}} \leq 0$. H2: at least 80% of not-determined units have $G_{\mathrm{ref}} > 0$. H3: the mean gap over not-determined units exceeds the mean over determined units, with a 95% family bootstrap interval that excludes 0. H4: on cohort B, the same separation with at least 80% of cells of the predicted sign. The 80% is a tolerance fixed in advance, and what we would conclude under each outcome was written into the freeze.

Table 34: The pre-registered labels by determinacy status: cells (encoder, label), cells at or below the input-matched floor chosen on validation, the mean G_{ref} over cells, family units at or below the floor, and the same cells on cohort B (`scripts/prereg_latex.py`).

label	status	cells	at or below floor	mean G_{ref}	family units at or below	cohort B cells	at or below
backbone H-bond count	determined	7	7	-0.124	3 of 3	3	3
backbone H-bond count	not determined	20	10	-0.055	7 of 12	3	0
backbone H-bond count	not established	5	4	-0.130	–	2	2
χ_1 class	determined	2	2	-0.084	1 of 1	2	2
χ_1 class	not determined	30	13	$+0.009$	10 of 19	6	1
χ_2 class	determined	2	2	-0.184	1 of 1	2	2
χ_2 class	not determined	30	14	$+0.041$	10 of 19	6	1
Cα virtual dihedral (control)	determined	22	22	-0.320	–	5	5
Cα virtual dihedral (control)	not determined	10	0	$+0.237$	–	3	0
side-chain contacts (descriptive)	determined	2	0	$+0.015$	–	2	0
side-chain contacts (descriptive)	not established	30	8	$+0.031$	–	6	0

Result. H1, H3 and H4 confirm; H2 does not (Table 34). Every determined unit falls below its floor, 5 of 5 (11 of 11 cells, each with a protein-level bootstrap interval below zero), and the mean gap separates the two groups by $+0.136$ $[+0.081, +0.188]$. Of the 50 not-determined units, 23 pass the floor, short of the 80% that H2 requires. The reading written into the freeze for this outcome is that determinacy predicts a graded shift in the gap without fixing its sign, and that encoders in general do not infer the missing information beyond fixed descriptors of their input. On cohort B, with the six frozen cohort encoders and the two GCPNets added before scoring so that the rotamer labels have a determined side, the gap separates the groups by $+0.253$ and 20 of 22 primary cells have the predicted sign; the two exceptions are ProteinMPNN on the two rotamer labels, which it falls below although its input does not determine them. The positive control holds: all 22 coordinate readers fall below the floor on the Cα dihedral and no sequence-only encoder does. On the descriptive side-chain contact count the two GCPNets, whose input determines it, pass the floor by $+0.015$.

Outside the frozen rules, the not-determined units that pass the floor are the sequence and multimodal models on every label and ESM-IF1 and PottsMPNN on both rotamer labels, while the Cα-only and graph networks, ProteinMPNN, Foldseek and GVP stay below, so whether an encoder passes the floor on a label its input does not fix also depends on the encoder.

E.4 The demand results on downstream proteins never scored

Section 3.6 reports that composition repairs its erasure on held-out proteins as well. This subsection asks whether the composition demand and its selectivity hold on downstream proteins that no threshold was developed against. The manifests hold more rows than the stage-1 subsample every Section 3.3 result was produced on. Those rows were never extracted and never scored, and eight cohorts were frozen from them, 13,632 proteins in all, preferring rows the source dataset already designates as test (`scripts/downstream_holdout_freeze.py`). Solubility and PPI affinity have no unused rows and are excluded without substitution, so the cohorts cover two of the three tasks the effect rule names for amino-acid composition. Heads, LEACE maps, standardisers and descriptor residualisations are fitted on the released training split and applied once (`scripts/addback_holdout.py`, six encoders, four sham permutations). EC classification and remote homology carry 373 classes and 895 folds in their training splits and the two GO tasks multilabel vocabularies, so these four were scored separately with the multi-output head on the same

cohorts (`scripts/bighead_holdout.py`, six encoders, four shams and four variance-matched nulls, with the shared eraser of `scripts/erasure.py`).

Table 35: The four downstream hypotheses as the rules frozen before scoring decide them.

	prediction, fixed before the cohort was scored	result on the held-out cohort	verdict
D1	Erasing linear access to amino-acid composition costs the retrained head more than the sham erasure does, on membrane and soluble and on subcellular localisation.	membrane vs soluble +0.043 over 6 encoders, 6 positive, one-sided p=0.0156; subcellular localisation +0.026 over 6 encoders, 6 positive, one-sided p=0.0156	confirmed
D2	The named descriptor repairs its own erasure more than it repairs the sham, and the permuted copy and the other structural blocks do not.	membrane vs soluble R=+0.0375 (6/6 positive) against permuted +0.0006, other +0.0015; subcellular localisation R=+0.0259 (6/6 positive) against permuted -0.0029, other -0.0009	confirmed
D3	Where the released run named nothing, nothing is named here either: GO-BP stays below the magnitude floor and stability does not hold in sign across the panel.	stability +0.063 over 6 encoders, 6 positive, one-sided p=0.0156: names composition. GO-BP is multilabel and not scored in this pass, so its half is undecided; the decidable half fails	undecided
D4	Remote homology remains a bound: no property reaches the magnitude floor.	remote homology not scored	undecided

D1 and D2 both confirm (Table 35). D1 predicts that erasing composition costs the head more than the sham erasure does, and D2 that the named descriptor repairs its own erasure more than the sham while the permuted copy and the other structural blocks do not. The demand for amino-acid composition and its selectivity therefore hold on proteins no threshold was developed against, at effect sizes close to those of the released run. D3 is undecided, and its decidable half fails: on these proteins the effect rule names composition on stability, where the released run named nothing. On stability the sham arm scores above the unerased representation on 5 of 6 encoders, against 0 of 6 on each of the other three tasks, so the effect there is inflated at the sham end; and with six encoders a sign test cannot return a p below 0.0156, so any six of six satisfies the frozen wording, a weakness of the frozen rule. D4 is undecided because remote homology carries 906 folds over the released splits, which the head of this pass cannot fit in reasonable time, and because only one of the six properties was erased; the multi-output tasks below include it.

Against the variance-matched null. The null of Appendix D.4 was built after these verdicts were decided, so it is outside the freeze and leaves them unchanged. Table 36 recomputes both contrasts against each null on the held-out cohorts, over the fifteen encoders of the released pass for the four cohorts above (the freeze fixed six) and the six of the multi-output runner for the other four tasks.

	variance removed			effect		repair contrast R	
task	erased	sham	null	vs. sham	vs. null	vs. sham	vs. null
EC classification	0.130	0.052	0.129	+0.0101 (4/6)	−0.0460 (2/6)	−0.0022 (3/6)	−0.0226 (2/6)
fluorescence	0.180	0.146	0.180	+0.0472 (15/15)	+0.0131 (11/15)	+0.0464 (15/15)	+0.0371 (13/15)
GO-BP	0.116	0.055	0.115	+0.0147 (5/6)	+0.0024 (3/6)	+0.0099 (5/6)	+0.0078 (4/6)
GO-MF	0.114	0.055	0.114	+0.0091 (6/6)	−0.0039 (1/6)	+0.0036 (4/6)	+0.0045 (5/6)
membrane vs soluble	0.171	0.078	0.171	+0.0460 (15/15)	+0.0040 (7/15)	+0.0441 (15/15)	+0.0349 (15/15)
remote homology	0.137	0.051	0.136	+0.0024 (3/6)	−0.0349 (1/6)	+0.0111 (4/6)	+0.0048 (3/6)
stability	0.216	0.072	0.215	+0.0715 (13/15)	−0.0138 (8/15)	+0.0615 (14/15)	+0.0447 (12/15)
subcellular localisation	0.162	0.085	0.162	+0.0261 (15/15)	−0.0276 (6/15)	+0.0255 (15/15)	+0.0212 (12/15)

Table 36: Both contrasts against both nulls, on the frozen held-out cohorts. The counts are the encoders on which the contrast is positive. EC classification, GO-BP, GO-MF and remote homology use the six-encoder panel of the multi-output runner; fluorescence, membrane vs soluble, stability and subcellular localisation use the fifteen encoders of the held-out pass. The property is amino-acid composition throughout.

The erasure effect loses its consistency in sign. Against the sham it is positive on every encoder of every task the standard effect rule names; against the matched null it is positive on 6 to 11 of 15, which is what chance produces ($+0.0040$ on 7 of 15 on membrane and soluble, -0.0276 on 6 of 15 on subcellular localisation). The repair contrast keeps its sign: $+0.0349$ on 15 of 15 encoders for membrane and soluble, $+0.0212$ on 12 of 15 for subcellular localisation, $+0.0371$ on 13 of 15 for fluorescence and $+0.0447$ on 12 of 15 for stability, where the level contrast is -0.0138 on 8 of 15. Under the paired protein bootstrap below the interval for R_{var} excludes zero on 9 to 13 encoders of 15, and the interval for Δ_{var} on 5 to 7 (`confirmation_downstream/holdout_bootstrap_matched_summary.csv`). The repair test therefore detects the stability demand behind D3's failed half, and the level contrast cannot attribute it; for this reason the paper names demands by repair.

The four multi-output tasks. None of EC classification, GO-BP, GO-MF and remote homology reaches the magnitude floor of 0.02 against either null (Table 36). EC classification and remote homology are null results of the released run, and GO-BP, where the released run named nothing, stays below the floor ($+0.0147$ against the sham). On GO-MF the released run named amino-acid composition at $+0.020$, exactly at the floor; on held-out proteins the effect is $+0.0091$ against the sham, positive on all six encoders and less than half the magnitude required, and -0.0039 against the matched null. GO-MF composition is the weakest of the eight pairs the effect rule names: the encoder-restricted rule of Appendix D.5 removes it, the family weighting of Appendix D.11 loses it, and the held-out cohort removes it as well. We therefore withdraw it. Under the repair rule GO-MF composition passes the sign test on the released panel below the floor ($+0.019$), and on these proteins its repair contrast against the matched null is $+0.0045$ on 5 of 6 encoders. On these four tasks the sham arm scores above the unerased representation on four to six of six encoders (ten of fifteen on stability, and none or one of fifteen on the other held-out tasks), following the output dimension: the LEACE map re-bases the representation onto an orthonormal basis, which improves the conditioning where the head is most overparameterised. The matched null, which removes as much as the real erasure, is above the unerased arm far less often (zero of six on remote homology against five for the sham, one against five on GO-MF, two against four on GO-BP), so the sham's advantage comes from the re-basing together with the small amount it removes. On EC classification every restricted arm scores above the unerased one, and only the ordering among the arms carries information.

Uncertainty over proteins. The seed spread of a downstream score measures training instability conditional on the proteins, and intervals over encoders or families measure variation across the panel; neither resamples proteins. The held-out run stored the per-protein prediction of every arm and every descriptor, so the evaluation proteins can be resampled with the same indices for every arm, descriptor and encoder of a task, and the difference of differences that defines R is computed on the same proteins throughout (`scripts/holdout_bootstrap.py`, 1,000 replicates).

Table 37: The held-out effect Δ_{sham} and repair contrast R_{sham} with intervals over proteins. **Intervals excluding 0** counts the encoders whose own 95% interval lies above zero. The last column is the panel mean of R_{sham} for the permuted copy of the descriptor and for the other structural blocks. The same intervals against the variance-matched null are given in the text above.

task	proteins	effect	intervals excl. 0	R named	intervals excl. 0	R permuted / other
fluorescence	2000	$+0.050$	6 / 6	$+0.0445$	6 / 6	-0.0006 / -0.0002
membrane vs soluble	1692	$+0.043$	6 / 6	$+0.0375$	6 / 6	$+0.0006$ / $+0.0015$
stability	2000	$+0.063$	4 / 6	$+0.0491$	4 / 6	-0.0136 / $+0.0002$
subcellular localisation	2000	$+0.026$	5 / 6	$+0.0259$	6 / 6	-0.0029 / -0.0009

On membrane and soluble and on fluorescence every encoder's interval for Δ_{sham} and for R_{sham} excludes zero (Table 37); on subcellular localisation every encoder's R_{sham} interval excludes zero and five of six effects do. The two controls stay within 0.0015 of zero on those three tasks. Stability is the exception, at 4 of 6 on both. The cohorts are UniProt accessions and GFP variants with no homology grouping in the manifest, so resampling by protein family is not possible, and the bootstrap conditions on the trained representations and fitted heads, so it leaves out the uncertainty from encoder pretraining and from the training set.

F TWO RESIDUE-LEVEL TASKS, AND WHAT THE POOLED READOUT HIDES

Table 38: Residue-level tasks. Group mean AUPRC and Spearman correlation ρ with pooled-task standing use all scored encoders. S: sequence-only; C: coordinate-reading. Determinacy refers to the supplied chain coordinates.

Task	Encoders (S/C)	S AUPRC	C AUPRC	ρ	Determined
Contact pairs	32 (10/22)	0.247	0.263	0.22	Yes
Interface residues	28 (10/18)	0.409	0.355	0.73	No

Section 3.2 reads pooled protein-level representations on all ten tasks, because the two tasks that used a residue-level or residue-pair readout are the two removed as unusable. A coordinate encoder is built to carry residue-resolved geometry, which a pooled readout cannot show, so Section 3.4 compares the pooled ranking with two residue-resolved tasks built from data already in the benchmark or rebuilt from a public set.

Contact pairs. The probing dataset carries pair labels, sampled at 512 pairs per domain with a minimum sequence separation of six, and residue-level embeddings are cached for every encoder. A contact is a CA–CA distance below 8 Å, which is 2.83% of the sampled pairs at the same rate in all three splits. The readout is a residue projection shared by both positions followed by a head on the sum, the absolute difference and the sequence separation, trained on the released superfamily-disjoint training split and scored by AUPRC (`scripts/t8_contact_task.py`). There are no floors, because all four coordinate floors are built from CA distances and each therefore contains the label.

Table 39: The two residue-resolved tasks per encoder, two-layer readout: contact pairs (AUPRC, base rate 0.028, and 0.021 for sequence separation alone) and interface residues on the 353 MaSIF-site test chains (AUPRC, mean of three seeds, base rate 0.177), each with the encoder's rank among those scored on the task, next to its mean rank among the 33 encoders over the ten pooled tasks. Encoders are ordered by that mean rank. Contact prediction has no floors, because every coordinate floor contains the label; the interface floors go through the identical probe (`scripts/residue_readout_report.py`).

			contact pairs		interface residues	
encoder	reads coordinates	pooled rank	AUPRC	rank	AUPRC	rank
ProtT5-XL	no	4.3	0.367	7	0.491	4
Ankh	no	4.5	0.367	8	0.432	7
ESM2-3B	no	5.0	0.192	19	0.448	6
ESM-1b	no	5.5	0.206	17	0.395	10
ESM3	yes	6.5	0.419	5	0.702	1
SaProt	yes	7.3	0.413	6	–	
ProstT5	no	7.4	0.533	2	0.559	3
ESM2-650M	no	7.8	0.203	18	0.449	5
ESM2-150M	no	8.7	0.191	20	0.391	12
ProtBert	no	9.9	0.136	28	0.353	16
ESM2-8M	no	11.1	0.125	30	0.259	24
ESM2-35M	no	11.2	0.151	26	0.315	20
ESM-IF1	yes	14.2	0.541	1	0.573	2
PottsMPNN-pre	yes	17.7	0.229	16	0.400	9
PottsMPNN-sol	yes	18.2	0.242	14	0.395	11
GVP	yes	18.3	0.129	29	0.376	13
PottsMPNN-ft	yes	18.9	0.232	15	0.403	8
GCPNet-pre	yes	21.4	0.183	21	–	
SchNet	yes	21.4	0.318	11	0.358	15
TFN-pre	yes	21.5	0.441	4	0.266	23
CDConv	yes	21.9	0.164	24	0.341	18
GearNet	yes	21.9	0.346	9	0.238	27
MACE	yes	22.6	0.156	25	0.341	19
GCPNet	yes	23.4	0.164	23	–	
IPA	yes	23.6	0.090	31	0.218	28
DimeNet	yes	24.2	0.071	32	0.287	21
PottsMPNN	yes	24.4	0.286	12	0.364	14
ProteinMPNN	yes	26.7	0.244	13	0.343	17
Foldseek	yes	26.9	0.146	27	–	
TFN	yes	27.3	0.489	3	0.250	26

Table 39 (continued)

encoder	reads coordinates	pooled rank	contact pairs AUPRC	contact pairs rank	interface residues AUPRC	interface residues rank
GearNet-pre	yes	29.1	0.322	10	0.272	22
IPA-pre	yes	30.9	0.169	22	0.256	25
`ref_length`	floor		–		0.180	
`ref_position`	floor		–		0.185	
`ref_aa`	floor		–		0.209	
`ref_seq_window`	floor		–		0.218	
`ref_seq_expert`	floor		–		0.231	
`ref_seq_global`	floor		–		0.219	
`ref_coords`	floor		–		0.310	
`ref_pairhist`	floor		–		0.334	
`ref_backbone`	floor		–		0.334	
`ref_global`	floor		–		0.350	

The group means do not separate (Table 39): coordinate-reading encoders average 0.263 over 22 and sequence-only 0.247 over 10, a difference a one-sided Mann-Whitney does not resolve ($p = 0.35$). The ordering changes. The correlation between contact AUPRC and standing on the ten pooled tasks is $\rho = +0.22$ ($p = 0.22$) over the 32 encoders scored on both, so the pooled ranking does not transfer, and the direction of the movement follows the input audit. The seven encoders that gain most are all coordinate-reading: TFN with random weights, near the bottom of the pooled suite, rises 24.3 places to third, pretrained GearNet 19.1, pretrained TFN 17.5, ProteinMPNN 13.7, ESM-IF1 13.2 into first, random GearNet 12.9 and PottsMPNN 12.4. The five that fall most are all sequence-only: ESM-2 at 8M falls 18.9 places, ProtBert-BFD 18.1, ESM-2 at 35M 14.8, ESM-2 at 3B 14.0 and ESM-1b 11.5. This is one task on one dataset and the two groups overlap, so the result restricts the conclusion of Section 3.2 to pooled readouts.

Interface residues. The contact label is a function of the chain's CA coordinates, which the coordinate encoders read. Does the pooled ranking transfer to a residue-level label the chain alone does not determine? Interface-site prediction was one of the four tasks removed as unusable, because the residue labels it was scored on mark chain membership (Appendix I). We rebuild it from the chains of MaSIF-site (Gainza et al., 2020), as distributed by ProteinWorkshop, with a label defined from geometry (`scripts/masif_build_labels.py`): a residue of the target chain is an interface residue if any of its heavy atoms lies within 5 Å of a heavy atom of another chain in the same file, and the encoders see the target chain alone. With the cleaning rules of the probing dataset, the 3,362 listed entries give 3,327 single target chains (5 have no structure file and 30 name several chains as the target), split into 2,677 training, 297 validation and 353 test chains from the MaSIF-site lists. Interface residues are 25.9% of the training residues and 17.7% of the test residues, and 2 of the 353 test chains have a training chain at 30% identity or more over half their length, none at 50% (MMseqs2). The readouts are those of the probing suite, trained on the training chains with early stopping on validation and scored on the test chains, and the ten floors go through the identical probes (`scripts/masif_interface_probe.py`). The panel is 28 of the 33 encoders: SaProt, Foldseek 3Di and both GCPNets read a structure file, which this task does not provide, and FoldVision has no residue states.

The result differs from the contact task (Table 39). The pooled ranking transfers: the correlation between interface AUPRC and standing on the ten pooled tasks is $\rho = +0.73$ ($p = 9 \times 10^{-6}$) over the 28 encoders, compared with $+0.22$ on contacts. The sequence-only group averages 0.409 and the coordinate-reading group 0.355, and the three highest scores belong to encoders pretrained at scale, ESM3 (0.702), ESM-IF1 (0.573) and ProstT5 (0.559). Every sequence-only encoder clears the best sequence floor (0.231), and 8 of the 18 coordinate encoders clear the best of all ten floors (`ref_global`, 0.350); the ten below it are the randomly initialised networks, CDConv, ProteinMPNN and the pretrained GearNet, IPA and TFN. The partner chain is absent, so the interface label is not a function of the chain the encoder sees, while the contact label is a function of the chain's CA coordinates. Of the two residue-resolved tasks, the coordinate encoders gain ranks on the one whose label their input determines, which matches the pattern of the probing comparison. Two tasks are too few to establish this as a rule.

G Extended related work

This section extends the related work of the main text: the encoders and benchmarks the panel draws on, probing studies of protein representations, and the work on probe baselines and erasure that the floors and the variance-matched null build on.

Protein Encoders Protein representation learning has progressed along three main directions. *Sequence-based encoders* treat amino acid sequences as language and use masked language modeling on large protein corpora to learn representations that capture evolutionary and structural information (Rives et al., 2019; Rao et al., 2021; Lin et al., 2023; Elnaggar et al., 2022; 2023; Brandes et al., 2022). *Structure-based encoders* operate directly on 3D coordinates using geometric message passing, equivariant transformations, or continuous spatial features to encode residue interactions and protein geometry (Zhang et al., 2022; Jing et al., 2020; Hermosilla & Ropinski, 2022; Fan et al., 2023; Morehead & Cheng, 2024). *Multimodal encoders* model sequence and structure jointly within one pretraining framework (Su et al., 2023; Hayes et al., 2025; Wang et al., 2024; Frolova et al., 2024; Su et al., 2025).

Protein Benchmarks Protein encoders are evaluated on benchmarks covering variant effect prediction, functional annotation, structural prediction, and related downstream tasks in both sequence-based and structure-based settings (Rao et al., 2019; Jankauskaitė et al., 2018; Dallago et al., 2021; Townshend et al., 2020; Kucera et al., 2023; Notin et al., 2023; Ye et al., 2024; Gao et al., 2025; Tan et al., 2025). TAPE established an early standard for evaluating sequence-based protein language models on downstream tasks (Rao et al., 2019), and later benchmarks broadened this scope to mutational effects, function prediction, and structure-aware evaluation. ProteinGym and PEER assess variant effect prediction and broader downstream utility (Notin et al., 2023; Xu et al., 2022), and ProteinInvBench emphasizes inverse folding and structure-conditioned evaluation (Gao et al., 2023). ProteinBench and PFMBench provide standardized comparative evaluations across diverse tasks and model classes (Ye et al., 2024; Gao et al., 2025). GeomMotif evaluates geometric preservation in protein generation (Strashnov et al., 2026).

Representation probing and interpretability Probing methods test whether specific properties are recoverable from frozen representations (Belinkov, 2021; Hewitt & Manning, 2019). In protein representation learning, early analyses showed that protein language models encode structural signals such as secondary structure and residue contacts, often in a layer-dependent manner (Vig et al., 2020). Later work broadened this analysis through attribution and information-theoretic approaches that relate learned representations to biochemical and structural properties (Zhang et al., 2024; Detlefsen et al., 2022). More recently, sparse feature discovery has exposed interpretable biological concepts in protein language models (Simon & Zou, 2025; Gujral et al., 2025; Adams et al., 2025; Garcia & Ansuini, 2025), while emerging work has moved from feature identification toward geometric, circuit-level, and interventional analyses (Tsui et al., 2026; Setlur et al., 2026). Complementary studies have examined how pretrained protein representations transfer across downstream tasks and how accessible different forms of biological information remain (Li et al., 2024; Rahimimovassagh et al., 2026). Together, these studies establish that protein encoders contain biologically meaningful information, but typically focus on a restricted set of properties, a particular model family, or a specific interpretability mechanism rather than systematically comparing many properties across heterogeneous encoders against input-matched baselines (Prabakaran & Bromberg, 2026). Our work complements this literature with a unified, labelwise analysis of representation accessibility across protein encoder families.

What a probe licenses, and interventional alternatives Outside protein work it is established that a probe score reflects the probe as well as the representation, and our determinacy argument is the protein-specific form of this point. Hewitt & Liang (2019) introduced control tasks and selectivity, arguing that a probe score needs a baseline of equal capacity before it can be interpreted. Pimentel et al. (2020) showed that under an unconstrained probe family the quantity being estimated is the mutual information between the label and the representation, which the data processing inequality bounds above by the mutual information with the input, with equality only when the representation is a sufficient statistic for the label; under such a probe family, representations differ in how easily the label can be extracted. Voita & Titov (2020) proposed description length so that ease of extraction is reported alongside accuracy. Our floors are a control of the same kind, specialized to labels that the model's own input determines: when $y = f(x)$ the appropriate baseline is a fixed descriptor of

x, and the comparison measures how accessible x stays in a representation. To our knowledge this determinacy condition has not been checked label by label for a protein probing suite, and eleven of the fifteen labels in our comparison satisfy it.

The second half of our work follows the interventional line. Ravfogel et al. (2020) removed a concept by iterated nullspace projection, and Belrose et al. (2023) gave the closed-form least-damaging linear erasure we use. Elazar et al. (2021) argued that the behavioural consequence of removing a property identifies whether a model uses it, and their study already pairs removal with a selectivity control of the kind we use in Appendix D.8; adding a descriptor back is an established control. We contribute the matched reference an erasure needs: a variance-matched null, which removes the same share of the representation's variance as the erasure, and the repair contrast R_{var}, which compares the gain from adding the property back after its erasure with the gain after that null. Around them we calibrate the construction with a sham erasure of the permuted property, which replaces the usual null of a random subspace of the same rank (Appendix D.4), a sensitivity ladder over ranks read at the rank the properties occupy, a confidence-bound decision per task, and a nonlinear re-probe that states what level of access the construction speaks about. For the released latent-space atlas we use debiased linear CKA (Kornblith et al., 2019) to compare encoders on the full embeddings, so that the comparison does not depend on a 2D projection.

H TASKS AND PROTOCOLS

This section defines the probing suite, its protocol and floors, the determinacy of each label, and the downstream tasks.

H.1 THE PROBING SUITE

Which structural and sequence properties does the probing of Section 3.1 measure? The suite has 21 probes in ten protein property groups (PPGs, I1–I10), each targeting one kind of structural, sequential or biochemical information at the residue, pair or protein scale (Table 40). Fifteen probes in nine PPGs enter the floor comparison of Section 3.1 and Figure 2. The other six are excluded by their definition, independent of their scores: ϕ and ψ were fitted both as 36-class bins and in a circular form, and the circular form is reported because it does not depend on a bin width; radius of gyration, asphericity and contact order were fitted raw and length-residualised, and the residualised form is reported because chain length alone predicts the raw one; and sequence recovery, the one probe of I10, is uninformative for any source that receives residue identity, since seven encoders and two floors score exactly 1.000 on it.

Table 40: Every probe in the suite, its protein property group (PPG) and its status. Fifteen enter the floor comparison; the other six are excluded for the stated reason, which does not depend on their result. *Determined* means the label is a closed-form function of the encoder's input; for two determined labels, contact density and contact order, the floor descriptor contains the label by construction. *Undetermined* means the label is not a function of the encoder's input, and *not established* means we can neither exhibit such a function nor rule one out (Appendix H.4). PPGs: I1 local conformation, I2 local curvature, I3 secondary structure, I4 solvent accessibility, I5 local flexibility, I6 burial and packing, I7 residue contacts, I8 global geometry, I9 fold architecture, I10 sequence.

probe	PPG	label status	in the comparison	note
backbone ϕ	I1	determined	yes	a closed-form function of the coordinates
backbone ψ	I1	determined	yes	a closed-form function of the coordinates
local curvature	I2	determined	yes	a closed-form function of the coordinates
DSSP-3	I3	determined	yes	a closed-form function of the coordinates
DSSP-8	I3	determined	yes	a closed-form function of the coordinates
RSA	I4	undetermined	yes	not a function of the encoder's input
RSA bins	I4	undetermined	yes	not a function of the encoder's input
B-factor	I5	undetermined	yes	not a function of the encoder's input
core/surface	I6	determined	yes	a closed-form function of the coordinates
contact density	I6	determined	yes	the floor contains the label by construction
CA distance bins	I7	determined	yes	a closed-form function of the coordinates

Table 40 (continued)

probe	PPG	label status	in the comparison	note
contact order	I8	determined	yes	the floor contains the label by construction
radius of gyration	I8	determined	yes	a closed-form function of the coordinates
asphericity	I8	determined	yes	a closed-form function of the coordinates
fold architecture	I9	not established	yes	neither shown to be a function of the encoder's input nor shown not to be
asphericity (raw)	I8	—	no	raw form; the length-residualised form is reported
contact order (raw)	I8	—	no	raw form; the length-residualised form is reported
ϕ bins	I1	—	no	binned twin of backbone ϕ; the circular form is reported
ψ bins	I1	—	no	binned twin of backbone ψ; the circular form is reported
radius of gyration (raw)	I8	—	no	raw form; the length-residualised form is reported, since length alone predicts it
sequence recovery	I10	—	no	uninformative for any source given residue identity: 7 encoders and 2 floors score 1.000

What each PPG measures. Every target is an established descriptor from structural biology, computed from the domain structure (`probing_data/make_probing_data.py`). I1, *local conformation*, is the backbone dihedrals ϕ and ψ, the rotations about the N–Cα and Cα–C bonds, in circular form. I2, *local curvature*, is the angle between consecutive Cα–Cα vectors, the bending of the trace where the dihedrals describe torsion. I3, *secondary structure*, is the DSSP assignment in three classes (helix, strand, coil) and in the eight DSSP classes. I4, *solvent accessibility*, is the relative solvent accessibility regressed on $[0, 1]$ and binned at 0.2 and 0.5 into buried, intermediate and exposed. I5, *local flexibility*, is the Cα B-factor z-scored within its chain, masked where the chain's B-factors are constant or zero. I6, *burial and packing*, is core against surface (core: at least 20 virtual Cβ neighbours within 10 Å) and the number of residues within 8 Å in four bins (0–5, 6–10, 11–15, 16 or more). I7, *residue contacts*, is the Cα distance of sampled residue pairs in 16 bins of 2.5 Å over 0–40 Å. I8, *global geometry*, is relative contact order, radius of gyration and asphericity, each residualised on log chain length with the fit made on the training split. I9, *fold architecture*, is the CATH architecture, the second level of the hierarchy. I10, *sequence*, is amino-acid identity. Two readouts are fitted on every probe, a linear layer and an MLP-2, and the readout family is chosen on validation (Appendix H.2).

H.2 Probe protocol

This subsection gives the protocol behind every probe score of Section 3.1 and behind the dataset and readouts summarised in Section 2.

Extraction. For each domain, residue embeddings are extracted once from the frozen encoder and cached for every probe. A coordinate encoder is handed the stored N, CA, C, O and CB coordinates with the sequence, or a structure file written from them; a sequence-only encoder is handed the sequence alone. The atom-level audit of Appendix B.1 records which of the five atoms each encoder reads. Language models are run in batches of up to 24,000 tokens and every other encoder one domain at a time, and no domain is truncated: the dataset keeps domains of 40 to 600 residues. An encoder is extracted only after it passes the input audit, and every extraction checks that the residue embedding has one row per labelled residue and no non-finite value, and records the number of distinct rows and the numerical rank of the pooled embeddings.

Readouts. Residue-level probes read per-residue embeddings, the pair-level probe reads the 512 residue pairs sampled per domain with a minimum sequence separation of six, and protein-level probes read mean-pooled residue embeddings or the encoder's global embedding. The encoder is never fine-tuned. The linear readout is a closed-form ridge regression for the regression and angular probes, with the penalty chosen on validation from $\{10^{-2}, \ldots, 10^{4}\}$, and a single softmax layer for the classification probes, trained once (seed 0) with early stopping on validation loss. The nonlinear

readout is an MLP-2, two hidden layers of width 128, trained with AdamW (learning rate 10^{-3}, weight decay 10^{-4}, batch size 2048) for at most 30 epochs with early stopping on validation loss (patience 3), over seeds 0, 1 and 2; the pair-level probe uses a shared residue projection of width 128 followed by the same kind of head. Classification probes use a cross-entropy loss and are scored by balanced accuracy, angular probes a circular loss and circular correlation, and regression probes MSE and Pearson correlation; padding positions are masked.

Which readout is reported. The released sweep scores each source on the test split and takes the better of the two readout families there. The comparison of Section 3.1 fits each readout of every source once and scores it on both the validation and the test split, so that the readout family of encoders and floors alike, and the floor itself, are chosen on validation and the test split is read once (Appendix B.2); its test scores reproduce those of the released sweep to within 10^{-8}. Where a section quotes a count under the released choice, it says so, and Appendix B.1 compares the rules. Figure 2A draws, for each PPG, the mean over its probes with the readout family chosen on validation; every count and gap in the main text is computed per probe.

Data and seeds. Experiments use CATH 4.3 with two group-disjoint splits, by superfamily (15,565 / 1,948 / 1,948 domains) and by topology (15,378 / 2,159 / 1,924), with zero group leakage in either. Residue-level and pair-level probes are fitted on a fixed subsample of 3,000 training and 500 validation domains and scored on 1,000 test domains. All labels are precomputed and stored in HDF5 (`probing_data/labels.h5`, 19,461 domains, 2,823,525 residues). The MLP-2 is evaluated with three seeds, which set probe initialisation and data order while the encoder weights stay fixed; results are the mean across seeds, with the seed spread in Appendix J. The ten non-learned floors go through the same probes, splits, readouts and seeds, and they serve as the control that a shuffled-label task provides in other probing studies. The software environment of every result is recorded with the release.

H.3 The non-learned floors, in full

Section 3.1 reads every probe score against ten non-learned floors. A floor is a fixed function of the input an encoder receives, passed through the identical probe, split, readout and seeds. Table 41 gives each one completely: the number of features the probe is handed, which atoms it draws on, the window, and what the features are.

Why a distance-only descriptor reaches 0.9998 on ψ. The backbone floors contain no angle, only the pairwise distances among the $5(2W+1)$ stored atoms of a $\pm W$ window and one chirality sign per offset. Distances plus a chirality sign determine the configuration of those atoms up to a rigid motion, and a dihedral is invariant under rigid motion, so ϕ and ψ, whose atoms all lie inside a ± 1 window, are functions of what the descriptor holds; on a determined label a floor that recovers the label confirms the determinacy argument. Removing the window (`ref_backbone_w0`, $W = 0$, only the residue's own five atoms) lowers backbone ϕ from 1.000 to 0.887, DSSP-8 from 0.751 to 0.453, DSSP-3 from 0.931 to 0.859 and local curvature from 1.000 to 0.946, as it should if the floor reaches them through the window rather than holding them directly. ψ stays at 0.999 because the peptide group is planar: the C=O direction of residue i fixes where N($i+1$) sits, so N, CA, C and O of the residue alone determine ψ. A floor measures what the stored atoms determine, which here is more than the four atoms in the definition of ψ suggest.

H.4 Determined, undetermined, and not established, label by label

The reading of a probe score in Section 3.1 depends on whether the stored atoms determine the label, and eleven of the fifteen probed labels are determined. Determinacy is separate from how predictable a label is. A label is *determined* when it is a closed-form function of the stored atom set: the same atoms always give the same label, with no reference to anything outside the file. A label is *undetermined* when two proteins with the same stored atoms can carry different labels, because the label is fixed by something the file does not contain. A label's determinacy is *not established* when we can exhibit neither the function nor two such proteins.

The eleven determined labels are checked against the stored data (Table 42). Eight are recomputed from the stored coordinates of 400 randomly drawn domains (the CA distances on 4 of them) with the derivation shown and compared with the stored label: the largest error is 5.2×10^{-5} rad on the

Table 41: The non-learned floors. Width is the number of features at residue level. No floor contains a precomputed angle, a local frame, or any quantity fitted on a label: every coordinate enters as an interatomic distance or as one chirality sign per residue offset.

floor	width	atoms	window	what it contains	why it was added
`ref_length`	2	–	–	log length and length/600	trivial baseline
`ref_position`	5	–	–	residue index, distance to the nearer terminus and its log, both scaled by length, log length	trivial baseline
`ref_aa`	20	–	own residue	the residue's amino-acid one-hot; at protein level, the composition	an encoder handed residue identity gets no credit for recovering it
`ref_seq_window`	357	–	± 8	amino-acid one-hot per offset with an out-of-chain flag; no property scale	context-matches the sequence side to the coordinate floors, which see a neighbourhood
`ref_seq_expert`	522	–	± 8, smoothed $3/7/15$	the one-hot window, seven published residue scales over the same window, their means over three window widths, chain composition, position	the sequence side needed a descriptor of the effort ref_global received
`ref_seq_global`	937	–	± 8, smoothed to 61	ref_seq_expert plus the chain's 400 dipeptide frequencies and two long-range scale averages	the non-local sequence analogue of ref_global
`ref_coords`	29	CA	16 nearest, five shells	sorted 16 nearest CA distances, chain-neighbour distances, contact counts at 5, 7, 9, 11, 13 Å	first coordinate floor; CA only, which understates the input
`ref_backbone`	326	N, CA, C, O, CB	± 2	all pairwise distances among the 5×5 atoms of the window, one chirality sign per offset, plus the ref_coords kNN and shell block. No angle is precomputed	encoders receive N, CA, C, O and CB, so a CA-only floor is too weak
`ref_global`	336	N, CA, C, O, CB	± 2 plus whole chain	ref_backbone plus the residue's distance and rank from the centroid and the leading eigenvector components of the 8 Å contact map	ref_backbone is local, so non-local probes still favoured the encoders
`ref_pairhist`	335	CA	whole chain	the joint histogram of (sequence separation, CA–CA distance) over all pairs, contact-map spectra at two cutoffs, and the residue's own row of the histogram	the non-local and fold-level probes needed the pair-distance distribution of the whole chain

backbone dihedrals, below 10^{-5} Å on the distances and zero on the two count-based labels. Contact order and the two DSSP labels are established by derivation: contact order is a function of the contact map, and DSSP computes its hydrogen-bond energies from the stored N, CA, C and O atoms.

Table 42: The numerical check of the eleven determined labels. Each of the first eight is recomputed from the stored coordinates with the derivation shown and compared with the stored label; the last column is the largest absolute difference over the domains checked, 400 drawn at random from the 19,461 of `probing_data/labels.h5`, and for CA distance bins the 2,048 stored pairs of 4 of them. The last two rows are established by derivation (`scripts/check_label_determinacy.py`, `probing_results/label_determinacy.csv`).

label	derivation from the stored atoms	domains	largest error
backbone ϕ	dihedral of C($i-1$), N(i), CA(i), C(i)	400	5.2×10^{-5} rad
backbone ψ	dihedral of N(i), CA(i), C(i), N($i+1$)	400	4.9×10^{-5} rad
local curvature	angle between consecutive CA–CA vectors	400	1.3×10^{-5} rad
contact density	count of CA within 8 Å	400	0 (count)
core/surface	CB neighbours within 10 Å, threshold 20	400	0 (class label)
radius of gyration	square root of the trace of the CA gyration tensor	400	1.9×10^{-6} Å
asphericity	gyration-tensor invariant	400	2.0×10^{-7} (unitless)
CA distance bins	the CA–CA distance itself	4 (2,048 pairs)	9.3×10^{-6} Å
contact order	mean sequence separation of contacts over length; derived from the contact map	–	not recomputed
DSSP-3 / DSSP-8	DSSP hydrogen-bond energies use N, CA, C, O; mkdssp reads the same atoms	–	not recomputed

Relative solvent accessibility is predictable and still undetermined: a fixed descriptor of the backbone reaches 0.900 on it, but RSA is the accessible surface of the full side chain, and two residues with

identical backbones and different side chains have identical descriptors and different RSA. The three encoders that read the structure file itself (both GCPNets and FoldVision) also receive the side-chain atoms, so for them RSA is determined, and the margin on RSA is read from the other encoders, where ESM3 leads by $+0.065$. B-factor is a crystallographic refinement quantity, which a conformation does not fix: the same fold refined twice carries different values. Fold architecture describes the arrangement of secondary structures, which the coordinates fix, but the CATH assignment also draws on evolutionary and functional evidence the file does not hold; we have neither written a function of the stored atoms that reproduces it nor found two domains that share an atom set and sit in different architectures, so its determinacy is not established, and the encoders beat every floor we built, including the structural alignment of Appendix B.8, on a label we cannot show their input determines.

The three statuses carry different readings. On a determined label an encoder scoring below the floor leaves the label less accessible to these readouts than a fixed function of its own input does; whether the representation still holds the label in a form a larger head could reach is outside this measurement. On an undetermined label no function of the stored atoms returns the label, and a score above the floor says the representation supports a better readout of the label than the floors do: the encoder and its readout are together a function of the input, so non-determinacy rules out exact reconstruction and leaves useful prediction possible. On a label whose determinacy is not established neither reading is available, and the floor score is the graded quantity Table 43 reports next to the flag.

Table 43: Every probe in the comparison with the argument for its status, the floor it is read against, and the best encoder. The determined labels are those a closed-form function of the stored atoms produces; the undetermined ones are fixed by side-chain atoms the extraction does not store or by a crystallographic measurement; fold architecture is the one label whose determinacy we neither establish nor refute. The best encoder on RSA and RSA bins is GCPNet-pretrained, which reads side-chain atoms from the structure file; without the three encoders that do, the best scores are 0.965 and 0.871 (ESM3).

probe	status	floor	best encoder	what fixes the label, and whether the stored atoms fix it
backbone ϕ	determined	1.000	0.997	the dihedral of C($i-1$), N(i), CA(i), C(i): all four atoms are stored, so the angle is a function of them
backbone ψ	determined	1.000	0.997	the dihedral of N(i), CA(i), C(i), N($i+1$): all four atoms are stored
local curvature	determined	1.000	0.993	the angle between consecutive CA–CA vectors: CA is stored
DSSP-3	determined	0.931	0.949	DSSP hydrogen-bond energies from N, CA, C, O: all four atoms are stored; DSSP places H from them
DSSP-8	determined	0.751	0.783	the same, at eight states: the same
core/surface	determined	0.897	0.937	a CB neighbour count at a fixed radius: CB is stored or virtual-CB is built
contact density	determined	0.984	0.780	a CA neighbour count at a fixed radius: CA is stored
CA distance bins	determined	0.556	0.363	the binned CA–CA distance: CA is stored
contact order	determined	0.983	0.923	a function of the contact map: CA is stored
radius of gyration	determined	0.996	0.937	the second moment of the CA coordinates: CA is stored
asphericity	determined	0.958	0.880	an invariant of the gyration tensor: CA is stored
RSA	undetermined	0.900	0.971	solvent accessibility of the full side chain, computed by freesasa: side-chain atoms beyond CB are not stored, and two residues with identical backbones and different side chains carry different RSA
RSA bins	undetermined	0.764	0.883	the same, binned: the same
B-factor	undetermined	0.587	0.660	a crystallographic refinement quantity: no function of the coordinates yields it; the same fold refined twice gives different values
fold architecture	not established	0.316	0.497	a CATH curator's assignment: the architecture level describes an arrangement of secondary structures, which the stored atoms fix, while the assignment also uses evidence that is not in the file; we exhibit neither the function nor a counterexample to it

The pattern the flag predicts holds: on the eleven determined labels the floor beats the best encoder on eight, and on the three undetermined labels and on fold architecture an encoder beats the floor. The count of eight is close to what the construction implies; the other four are the empirical content, because nothing in the construction requires an encoder to beat a floor on a label the floor is not shown to compute. The released rule, which holds every encoder to the floors of its input group and reads every score on test, gives 233 of 245 coordinate cells below the best of the ten floors and 0 of 110 sequence-only cells below the best sequence floor; Appendix B.1 compares it with the input-matched rule, whose count Section 3.1 reports.

H.5 THE DOWNSTREAM TASKS

This subsection defines the downstream tasks of Section 3.2 and the protocol that scores them. Of the fourteen tasks assembled, four are unusable as labelled and are excluded (Appendix I); the other ten are scored. All evaluations use frozen embeddings, so the score reflects the representation and involves no task-specific fine-tuning. Every encoder is evaluated on the ten scored tasks except FoldVision, which has no remote-homology extraction. Where the source dataset provides no structure, structures are predicted, so no task is skipped because of an encoder's input modality. Table 44 gives the prediction target, level, metric and source of each task, and Table 45 the structures and splits of the ten scored ones.

Table 44: Downstream tasks used to evaluate the transfer performance of frozen encoder representations. "Level" gives whether prediction is made for a protein, a residue or a residue pair. Protein-level tasks read pooled embeddings, and residue- and pair-level tasks read per-residue embeddings. T1, T5, T10 and T11 are excluded as unusable as labelled (Appendix I); residue-pair and residue-level readouts are tested on the rebuilt tasks of Appendices F and F.

Task	Name	Description	Level	Metric	Dataset
Structure Prediction (T1)					
T1	Contact prediction	Residue pairs with Cα $<$ 8Å	Pair	Top L/5 (Wang et al., 2017)	ProteinNet (AlQuraishi, 2019)
Fitness Landscape (T3–T5)					
T3	Fluorescence	GFP fluorescence prediction	Protein	Spearman ρ	TAPE (Rao et al., 2019)
T4	Stability	Thermostability prediction	Protein	Spearman ρ	TAPE (Rao et al., 2019)
T5	Mutation effect	Functional effect of mutations	Protein	Spearman ρ	ProteinGym (Notin et al., 2023)
Functional Annotation (T2, T6–T9)					
T2	Remote homology	Fold classification at low sequence identity	Protein	Accuracy	TAPE (Rao et al., 2019)
T6	EC classification	Enzyme classification (373 classes in the training split; 384 in the manifest)	Protein	Accuracy	ProteinWorkshop (Jamasb et al., 2024)
T7	GO-MF	Molecular function prediction (485 terms)	Protein	F_{max}	ProteinWorkshop (Jamasb et al., 2024)
T8	GO-BP	Biological process prediction (1,926 terms)	Protein	F_{max}	ProteinWorkshop (Jamasb et al., 2024)
T9	Solubility	Protein solubility prediction	Protein	AUROC	FLIP (Dallago et al., 2021)
Protein–Protein Interactions (T10–T12)					
T10	PPI sites	Interface residue prediction	Residue	AUPRC	MaSIF-site (Gainza et al., 2020)
T11	Metal binding	Metal-binding residue prediction	Residue	AUPRC	Metal3D (Dürr et al., 2023)
T12	PPI affinity	Binding affinity change ($\Delta\Delta G$) prediction	Protein	Spearman ρ	SKEMPI2 (Jankauskaitė et al., 2018)
Subcellular Localization (T13–T14)					
T13	Localization (10-class)	Subcellular localization prediction	Protein	Accuracy	DeepLoc2 (Thumuluri et al., 2022)
T14	Membrane/ soluble	Membrane versus soluble classification	Protein	AUROC	DeepLoc2 (Thumuluri et al., 2022)

Protocol. Representations are extracted once from each frozen encoder and cached (released on Hugging Face), and every task reads a pooled protein-level representation; for PPI affinity the feature of a pair is the concatenation of the two chains' embeddings in the order the dataset gives. Two heads are fitted to every cell, the same for every encoder (`extract/score2.py`): one linear layer, and an MLP-2 with two hidden layers of width 512, ReLU activations and dropout 0.1 after the first, both on features standardised on the training split and trained with AdamW (learning rate 10^{-3}, weight decay 10^{-4}, batch 1,024) for at most 60 epochs with early stopping at patience 5 on the validation loss. The input dimension varies from 20 to 2,560, and with it the number of parameters in each

Table 45: Structures and splits of the ten scored tasks: proteins (pairs for PPI affinity) used for training, validation and test, with the training rows available in parentheses where the training split is subsampled to the cap of 6,000. The subsample is fixed in the released manifest.

task	structures	train (available)	validation	test
EC classification	experimental (PDB)	6,000 (28,047)	1,500	3,000
GO-MF	experimental (PDB)	6,000 (23,104)	1,500	2,904
GO-BP	experimental (PDB)	6,000 (21,552)	1,500	2,903
remote homology	experimental (PDB)	6,000 (8,122)	506	480
subcellular localisation	predicted (AlphaFold DB)	6,000 (17,984)	1,500	2,473
membrane vs soluble	predicted (AlphaFold DB)	6,000 (14,968)	1,500	3,000
stability	predicted (ESMFold)	6,000 (46,813)	1,500	3,000
fluorescence	predicted (ESMFold)	6,000 (21,446)	1,500	3,000
solubility	predicted (ESMFold)	5,417	1,363	1,329
PPI affinity	predicted (ESMFold)	2,352 pairs	194 pairs	326 pairs

head's first layer, so Appendix C.3 repeats the comparison with every representation projected to 116 dimensions. Each cell is evaluated with three seeds; the head is selected by mean validation loss across seeds, and the reported cell is the mean over the three seeds of that head, with the seed spread released alongside (`scores2_selected.csv`). Metrics follow the task: Spearman ρ for stability, fluorescence and PPI affinity, AUROC for solubility and membrane vs soluble, accuracy for remote homology, EC classification and subcellular localisation, and protein-centric $F_{\max}$, the best F1 over a sweep of thresholds, for GO-MF and GO-BP.

H.6 Downstream results in detail

The paragraphs below give the per-task numbers behind Section 3.2. Training is capped at 6,000 proteins on the eight tasks whose training split is larger (Table 45); solubility and PPI affinity are used whole. The MLP-2 head is selected in 270 of the 329 cells and the linear head in 59, of which 32 are EC classification. The class and term counts that set the output size of each head are those of the training split: 373 EC classes, 895 remote-homology folds (906 over the three splits), 10 localisation classes, and 485 and 1,926 GO terms.

Over the whole panel, the sequence-only encoders lead on every task. Ten of the 33 encoders are measured not to read coordinates (`uses_coordinates` false): their pooled embedding stays fixed when the coordinates are replaced with those of another protein of the same length. The remaining 23 read coordinates. The group mean of the ten exceeds the group mean of the 23 on every task: EC classification 0.745 against 0.273, stability 0.614 against 0.260, GO-MF 0.508 against 0.273, subcellular localisation 0.664 against 0.476, fluorescence 0.541 against 0.358, remote homology 0.254 against 0.134, solubility 0.689 against 0.584, membrane versus soluble 0.903 against 0.800, GO-BP 0.376 against 0.286, and PPI affinity 0.250 against 0.240. Nine of these ten gaps are between 0.090 and 0.473, with a mean over the ten tasks of 0.186; the tenth, PPI affinity at 0.011, is a tie and is treated as one below. The declared classes give a similar ordering. The nine sequence-based encoders lead the group mean on eight of the ten tasks (EC 0.738 against 0.579 multimodal and 0.220 structure-based; stability 0.641 / 0.498 / 0.203; GO-MF 0.504 / 0.424 / 0.248; subcellular localisation 0.668 / 0.579 / 0.456; fluorescence 0.537 / 0.423 / 0.352; solubility 0.691 / 0.656 / 0.570; membrane versus soluble 0.903 / 0.870 / 0.787; GO-BP 0.374 / 0.343 / 0.276), and the five multimodal encoders lead on remote homology (0.282 / 0.241 / 0.115) and PPI affinity (0.337 / 0.256 / 0.212), the latter being the column discussed below. The structure-based class is last on all ten. Seven of the ten per-task maxima are held by a sequence-only encoder: Ankh-base on fluorescence (0.587), GO-BP (0.405) and GO-MF (0.565), ProtTrans-T5 on solubility (0.707) and subcellular localisation (0.705), one of ESM-2(3B), Ankh-base or ESM-1b on EC classification, and one of ESM-2(3B), ProtTrans-T5 or ESM-1b on membrane versus soluble. The holder of the last two is left unnamed, because a lead smaller than its own seed spread is not reported as a lead: on EC classification the top three scores are 0.8185, 0.8182 and 0.8115 against a seed standard deviation of 0.0057 on the leader, so the top-two margin of 0.0003 is a twentieth of the noise, and on membrane versus soluble they are 0.9135, 0.9133 and 0.9131 against 0.0070. By mean rank across the ten tasks, the first twelve places are the ten sequence-only encoders together with ESM3 (6.50) and SaProt (7.30), led by ProtTrans-T5 at 4.30

and Ankh-base at 4.45; the best encoder that reads coordinates and not residue identity is ESM-IF1 at mean rank 14.2.

The coordinate encoders that compete across tasks are the two that were also pretrained on sequence at scale. ESM3 holds two of the three per-task maxima that do not go to a sequence-only encoder, remote homology (0.403) and stability (0.777), which are +0.037 and +0.024 above the best sequence-only encoder on those tasks; on the other eight tasks it is below the best sequence-only encoder by 0.002 to 0.128. Across the ten tasks there are 20 cells in which a coordinate encoder exceeds the sequence-only group mean, and 10 of them belong to ESM3 or SaProt. Of the other ten, nine are on PPI affinity and the tenth is ESM-IF1 on remote homology (0.358). Without ESM3, SaProt and FoldVision, the mean of the remaining 20 coordinate encoders falls from 0.273 to 0.215 on EC classification, from 0.260 to 0.201 on stability and from 0.273 to 0.248 on GO-MF.

PPI affinity is the one column that does not separate the groups, and it is also the noisiest. Its test split is the smallest in the suite at 326 pairs, and it holds the third per-task maximum that does not go to a sequence-only encoder: IPA, a randomly initialised geometric network, scores 0.499, above every sequence-only encoder, the best of which is ESM-2(3B) at 0.425. Eleven of the 23 coordinate encoders exceed the sequence-only group mean on this task, against at most three on any other. The seed spread explains the column: its mean seed standard deviation is 0.101, against 0.011 over the other nine tasks together, and it holds the five largest cell spreads in the suite, up to 0.255 for ESM3. Scoring the same runs by the best of the three seeds on validation, in place of the seed mean, moves 14 cells of this column by more than 0.1 and changes which encoder holds the task; no task other than PPI affinity and stability has a single cell that moves by more than 0.046 under that substitution. Differences in this column below roughly 0.2 are not resolved by three seeds, and we draw no conclusion from it. Outside PPI affinity and stability the scores are stable across seeds: the median seed standard deviation over all 329 encoder cells is 0.0059, the per-task median is below 0.009 on eight of the ten tasks, and the largest single spread on those eight is 0.040.

Pretraining a geometric network moves its downstream scores in both directions. Ten of the 23 coordinate encoders carry random weights, so the panel contains matched random and pretrained versions of the same architecture. Averaged over the ten tasks, pretraining changes the score by +0.063 for GCPNet, +0.055 for PottsMPNN, +0.035 for TFN, −0.062 for GearNet and −0.112 for IPA. Within a pair the per-task effect is larger than the average in both directions: GCPNet gains 0.549 on stability and loses at most 0.027 on any other task, while IPA loses 0.343 on PPI affinity and GearNet loses 0.162 on EC classification. The random-weight subgroup exceeds the pretrained coordinate subgroup on fluorescence (0.368 against 0.350) and on PPI affinity (0.255 against 0.228).

The comparison is confounded by pretraining data and scale. The two groups differ in more than their input. Pretraining corpus and model scale differ with the grouping and neither is controlled in this comparison, so the gap cannot be attributed to the input modality alone; Appendix C.3 holds these confounds fixed, and Section 3.2 reports that among large pretrained encoders the two groups are not separated. Embedding width alone tracks the grouping closely: the ten sequence-only encoders span 320 to 2560 dimensions, while every coordinate network is between 20 and 512 except ESM3 (1536), SaProt (1280) and FoldVision (1024), and the two coordinate encoders that compete, ESM3 and SaProt, are the two widest of the 23.

Table 46 is the downstream matrix of Section 3.2: 33 encoders by ten scored tasks, each cell the mean over three seeds of the head selected on validation loss. The metric differs by task and is named in the column header, so scores are comparable within a column. A mean score across different metrics has no meaning, so the summary column is a mean rank. Encoders are grouped by the measured input audit, which records how each embedding responds to a coordinate jitter, a coordinate swap, a sequence shuffle and a rigid motion; the declared class is kept as a label. Ten encoders are measured not to read coordinates, 16 to read coordinates and residue identity, and 7 to read coordinates and not residue identity. One cell is empty, FoldVision on remote homology, for which no extraction exists. Table 47 gives the seed standard deviation of every cell next to the score: the median over the 329 encoder cells is 0.0059, and PPI affinity (median 0.0671, maximum 0.2545) and stability (median 0.0493, maximum 0.1210) carry almost all of the spread. Appendix J discusses the distribution.

Table 46: Downstream task scores on all ten scored tasks. Each cell is the mean over three seeds of the head (linear or MLP-2) selected on validation loss; the metric is named in the column header. Encoders are grouped by the *measured* input audit and sorted within a group by mean rank over the tasks they have a score for (column MR, rank 1 best). The best score in each column is bold. The one empty cell is FoldVision on remote homology, for which no extraction exists. The last row is the non-learned composition reference of Appendix C.1, scored through the identical protocol. Seed standard deviations are in Table 47.

Encoder	EC	R.Hom.	Subcell.	GO-MF	GO-BP	Solub.	Memb.	Fluor.	Stab.	PPI Aff.	MR
	Acc.	*Acc.*	*Acc.*	*F-max*	*F-max*	*AUROC*	*AUROC*	*Sp.*	*Sp.*	*Sp.*	
Measured to read sequence only (10 encoders)											
ProtT5-XL	0.803	0.238	**0.705**	0.540	0.388	**0.707**	0.913	0.544	0.697	0.345	4.3
Ankh	0.818	0.294	0.662	**0.565**	**0.405**	0.690	0.907	**0.587**	0.752	0.254	4.5
ESM2-3B	**0.819**	0.284	0.701	0.549	0.385	0.688	**0.913**	0.500	0.535	0.425	5.0
ESM-1b	0.811	0.274	0.682	0.534	0.390	0.690	0.913	0.534	0.690	0.303	5.5
ProstT5	0.807	0.365	0.628	0.544	0.392	0.669	0.902	0.579	0.373	0.199	7.4
ESM2-650M	0.791	0.246	0.690	0.530	0.385	0.706	0.904	0.518	0.578	0.161	7.8
ESM2-150M	0.757	0.251	0.668	0.485	0.366	0.691	0.900	0.508	0.654	0.196	8.7
ProtBert	0.590	0.148	0.650	0.426	0.347	0.705	0.899	0.518	0.595	0.331	9.9
ESM2-8M	0.546	0.198	0.611	0.428	0.339	0.657	0.885	0.570	0.640	0.249	11.1
ESM2-35M	0.709	0.239	0.644	0.476	0.363	0.687	0.895	0.553	0.627	0.042	11.2
Measured to read coordinates (23 encoders)											
ESM3	0.802	**0.403**	0.616	0.533	0.381	0.705	0.889	0.459	**0.777**	0.338	6.5
SaProt	0.687	0.298	0.659	0.486	0.369	0.680	0.902	0.427	0.715	0.439	7.3
ESM-IF1	0.714	0.358	0.505	0.419	0.329	0.642	0.831	0.408	0.235	0.144	14.2
FoldVision	0.471	–	0.529	0.317	0.296	0.639	0.865	0.295	0.466	0.265	15.6
PottsMPNN-pre	0.401	0.172	0.455	0.295	0.285	0.605	0.808	0.363	0.220	0.304	17.7
PottsMPNN-sol	0.385	0.190	0.472	0.290	0.285	0.569	0.792	0.363	0.276	0.331	18.2
GVP	0.190	0.105	0.475	0.251	0.285	0.598	0.804	0.392	0.212	0.357	18.3
PottsMPNN-ft	0.416	0.171	0.466	0.293	0.287	0.578	0.817	0.363	0.264	0.160	18.9
SchNet	0.148	0.054	0.483	0.238	0.279	0.595	0.795	0.378	0.284	0.140	21.4
GCPNet-pre	0.147	0.135	0.458	0.235	0.268	0.585	0.828	0.346	0.482	0.264	21.4
TFN-pre	0.155	0.050	0.497	0.242	0.277	0.548	0.826	0.356	0.323	0.190	21.5
CDConv	0.127	0.062	0.464	0.237	0.279	0.588	0.792	0.353	0.159	0.442	21.9
GearNet	0.198	0.047	0.488	0.236	0.274	0.613	0.788	0.446	0.232	0.150	21.9
MACE	0.145	0.076	0.481	0.241	0.281	0.580	0.778	0.392	0.261	0.071	22.6
GCPNet	0.103	0.107	0.485	0.230	0.274	0.582	0.832	0.343	-0.068	0.228	23.4
IPA	0.199	0.068	0.451	0.226	0.275	0.539	0.785	0.353	0.177	**0.499**	23.6
DimeNet	0.133	0.057	0.466	0.227	0.275	0.558	0.778	0.364	0.037	0.329	24.2
PottsMPNN	0.199	0.129	0.439	0.239	0.267	0.546	0.778	0.337	0.089	0.338	24.4
ProteinMPNN	0.204	0.149	0.420	0.236	0.268	0.544	0.775	0.328	0.110	0.121	26.7
Foldseek	0.240	0.153	0.392	0.229	0.259	0.537	0.710	0.270	0.280	0.124	26.9
TFN	0.110	0.048	0.455	0.233	0.273	0.555	0.780	0.321	0.348	-0.004	27.3
GearNet-pre	0.036	0.052	0.403	0.174	0.254	0.556	0.700	0.363	0.182	0.130	29.1
IPA-pre	0.057	0.066	0.380	0.179	0.253	0.496	0.741	0.210	-0.088	0.156	30.9
composition reference	0.086	0.071	0.470	0.227	0.274	0.588	0.779	0.362	0.090	0.091	–

I TASKS AND ANALYSES THE PAPER EVALUATED AND DOES NOT USE

Four tasks unusable as labelled. Four of the fourteen assembled tasks are excluded. PPI site prediction is labelled by chain membership, where the intended label is interface contact, so the residue index alone reaches AUPRC 0.971, and every encoder scored on it reached above 0.90 on a label it did not need the embedding to predict. Metal binding is labelled by proximity to any heteroatom and is 89.8% positive. Contact prediction was computed from CA-only coordinates indexed as though they were N, CA and C, so the contact map it scored differs from the contact map of the protein. Mutation effect was reported as a single Spearman correlation pooled over 176 assays, which mixes assays with incompatible sign and scale. Two of the four defects arose in this benchmark's assembly: the misindexed contact map came from the benchmark's own extraction and evaluation code, and the pooled Spearman was the metric that code computed for mutation effect. For PPI site prediction and metal binding the origin of the defect is not established. The ProteinWorkshop copies of MaSIF-site and Metal3D that the benchmark drew on store each structure with a per-residue chain index, and with the name of the target chain (MaSIF-site) or the heteroatoms (Metal3D), but no residue label. The step that produced the residue labels the benchmark scored is not among its files, so we cannot say whether the chain-membership labels and the any-heteroatom rule come from the

upstream release or from this benchmark's assembly. These four tasks are removed and not re-scored; recomputing mutation effect per assay is possible and is not done here. The remaining ten are rebuilt, with label definitions that match the intended label and manifests that enforce one common ID set per task with no protein in two splits.

Correlational estimates of demand. Can a correlation over encoders between probe scores and task scores show which properties a task uses? The paper measures demand by intervention (Section 3.3), and this paragraph gives the evidence that a correlation cannot replace it. The information alignment analysis of the earlier version estimates a task's demand for each PPG from the correlation, over encoders, between probe scores and task scores, and the encoder selection predictor ranks encoders for a held-out task from those estimates. The predictor chooses the neighbour tasks of a held-out task with that task's own scores, so the held-out scores enter the prediction. Without that selection it reaches a median Spearman of 0.674 over the 14 tasks of that version, which include the four excluded above, and two predictors that use no probing information reach 0.690 and 0.718. The task matrix of that analysis also held 72 exact zeros, placeholders for runs that had failed; with them treated as missing, which leaves 12 of the 14 tasks, the pooled ridge on probing profiles reaches 0.698 and the two predictors without probing information 0.759 and 0.735, so the comparison goes the same way.

The correlational form also has a limitation that no estimator removes: an encoder that decodes one property well decodes most of them well and scores well on every task. A partial rank correlation over the 33 encoders between each task's score and the readout of each property, adjusted for the encoder's general readout strength and its downstream standing, corroborates none of the eight pairs the standard effect rule names: no contrast between the demanded property and the strongest undemanded one has a bootstrap interval above zero, and three are below it (`scripts/demand_prediction_test_v2.py`; the table is released). A correlation over encoders therefore cannot separate what a task needs from what these representations carry, and no number in the paper comes from these analyses.

J SEED VARIABILITY

How much do the probe and downstream scores move between seeds, and is that spread small next to the differences Sections 3.1 and 3.2 interpret?

The probing suite. The linear readout is fitted once (Appendix H.2), so the spread quoted here is that of the MLP-2 readout over seeds 0, 1 and 2, on the headline metric of each probe. Over the 885 (encoder or floor, probe) cells of the 33 encoders and ten floors, the median seed standard deviation is 0.0029, the mean 0.0063 and the maximum 0.1036 (GearNet (pretrained) on asphericity). The variability is concentrated in the protein-level probes, where a single fit sees far fewer examples: the per-probe median is 0.0126 for contact order, 0.0108 for fold architecture, 0.0104 for asphericity and 0.0101 for radius of gyration, against 0.0003 for ϕ, 0.0010 for ψ and 0.0006 for RSA (Figure 7). The fold-architecture comparison of Section 3.1 is the one place where a difference of this size would change a conclusion, so it is run with 15 seeds and the rest of the suite with three. On the topology-disjoint split its per-cell seed standard deviation has median 0.0166 over the eleven shared classes and 0.0154 over all thirteen, with a maximum of 0.0399, and the interval of $[+0.057, +0.099]$ on the eleven shared classes (Appendix B.8) is computed from these 15 seeds.

The downstream tasks. Each cell of Table 46 is the mean over three seeds of the head selected on validation loss, and Table 47 gives the seed standard deviation of the same three fits; 329 of the 330 cells are scored, the missing one being FoldVision on remote homology. Over the 329 cells the seed standard deviation has median 0.0059, mean 0.0202 and maximum 0.2545, and 290 cells are below 0.05. Eight tasks have a per-task median below 0.009 and a maximum no larger than 0.040: GO-BP (median 0.0024, max 0.0059), GO-MF (0.0031, 0.0085), membrane and soluble (0.0036, 0.0082), subcellular localisation (0.0045, 0.0166), remote homology (0.0054, 0.0111), solubility (0.0067, 0.0330), EC classification (0.0074, 0.0400) and fluorescence (0.0080, 0.0222), an order of magnitude below the differences the main text draws conclusions from. PPI affinity (median 0.0671, with the five largest cell spreads of the suite) and stability (median 0.0493, maximum 0.1210) hold all

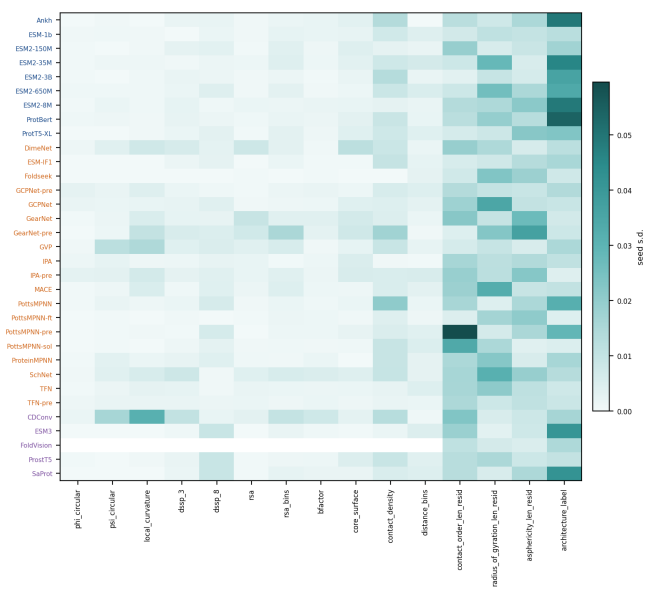


Figure 7: Seed standard deviation of the three-seed MLP-2 readout across (encoder, probe) cells. The mean probe score of every encoder with its seed spread is released as figures (`scripts/make_uncertainty_figures.py`).

16 cells above 0.1; differences below roughly 0.2 and 0.1 are not resolved there, and Appendix H.6 discusses the PPI affinity column.

Table 47: Seed standard deviation of every cell of Table 46, over the three seeds of the selected head. Median 0.0059 over the 329 encoder cells; PPI affinity (median 0.0671) and stability (0.0493) carry the spread.

Encoder	EC	R.Hom.	Subcell.	GO-MF	GO-BP	Solub.	Memb.	Fluor.	Stab.	PPI Aff.	MR
	Acc.	*Acc.*	*Acc.*	*F-max*	*F-max*	*AUROC*	*AUROC*	*Sp.*	*Sp.*	*Sp.*	
Measured to read sequence only (10 encoders)											
ProtT5-XL	0.0023	0.0064	0.0037	0.0060	0.0037	0.0150	0.0033	0.0105	0.0185	0.1143	4.3
Ankh	0.0040	0.0063	0.0069	0.0029	0.0047	0.0037	0.0030	0.0109	0.0431	0.0589	4.5
ESM2-3B	0.0057	0.0064	0.0074	0.0085	0.0058	0.0097	0.0070	0.0222	0.0624	0.1019	5.0
ESM-1b	0.0154	0.0080	0.0088	0.0027	0.0035	0.0014	0.0015	0.0100	0.0491	0.2155	5.5
ProstT5	0.0031	0.0111	0.0035	0.0035	0.0059	0.0012	0.0005	0.0071	0.0518	0.0671	7.4
ESM2-650M	0.0010	0.0042	0.0036	0.0050	0.0044	0.0063	0.0029	0.0087	0.0714	0.0633	7.8
ESM2-150M	0.0016	0.0044	0.0027	0.0038	0.0048	0.0050	0.0034	0.0111	0.0297	0.1567	8.7
ProtBert	0.0012	0.0049	0.0026	0.0045	0.0022	0.0023	0.0028	0.0039	0.0276	0.2058	9.9
ESM2-8M	0.0069	0.0042	0.0084	0.0064	0.0011	0.0069	0.0011	0.0023	0.0322	0.1993	11.1
ESM2-35M	0.0062	0.0049	0.0047	0.0051	0.0032	0.0109	0.0033	0.0073	0.0633	0.0458	11.2
Measured to read coordinates (23 encoders)											
ESM3	0.0102	0.0012	0.0028	0.0018	0.0047	0.0008	0.0043	0.0105	0.0317	0.2545	6.5
SaProt	0.0070	0.0087	0.0040	0.0058	0.0055	0.0074	0.0003	0.0081	0.0303	0.1654	7.3
ESM-IF1	0.0029	0.0032	0.0043	0.0016	0.0025	0.0067	0.0059	0.0047	0.1109	0.0662	14.2
FoldVision	0.0035	–	0.0079	0.0053	0.0028	0.0013	0.0057	0.0211	0.0182	0.0901	15.6
PottsMPNN-pre	0.0099	0.0032	0.0055	0.0025	0.0015	0.0135	0.0036	0.0097	0.0742	0.0351	17.7
PottsMPNN-sol	0.0036	0.0068	0.0114	0.0014	0.0024	0.0098	0.0059	0.0092	0.0935	0.0192	18.2
GVP	0.0167	0.0076	0.0057	0.0040	0.0015	0.0073	0.0061	0.0066	0.0096	0.0603	18.3
PottsMPNN-ft	0.0111	0.0080	0.0115	0.0024	0.0026	0.0145	0.0012	0.0029	0.0639	0.1566	18.9
SchNet	0.0114	0.0012	0.0015	0.0031	0.0032	0.0113	0.0082	0.0071	0.0925	0.2386	21.4
GCPNet-pre	0.0038	0.0042	0.0166	0.0025	0.0009	0.0248	0.0051	0.0055	0.0268	0.0388	21.4
TFN-pre	0.0368	0.0056	0.0137	0.0030	0.0023	0.0064	0.0057	0.0148	0.0773	0.0447	21.5
CDConv	0.0141	0.0064	0.0022	0.0022	0.0039	0.0131	0.0062	0.0014	0.0037	0.0513	21.9
GearNet	0.0034	0.0053	0.0039	0.0050	0.0020	0.0076	0.0050	0.0097	0.0777	0.1971	21.9
MACE	0.0255	0.0084	0.0045	0.0065	0.0029	0.0057	0.0027	0.0046	0.0922	0.0590	22.6
GCPNet	0.0240	0.0036	0.0077	0.0031	0.0018	0.0111	0.0062	0.0044	0.0316	0.0097	23.4
IPA	0.0074	0.0032	0.0085	0.0019	0.0018	0.0015	0.0015	0.0081	0.1210	0.1032	23.6
DimeNet	0.0069	0.0096	0.0012	0.0052	0.0006	0.0080	0.0022	0.0106	0.0376	0.2265	24.2
PottsMPNN	0.0084	0.0053	0.0041	0.0017	0.0021	0.0001	0.0048	0.0085	0.0872	0.0260	24.4

Table 47 (continued)

Encoder	EC	R.Hom.	Subcell.	GO-MF	GO-BP	Solub.	Memb.	Fluor.	Stab.	PPI Aff.	MR
	Acc.	*Acc.*	*Acc.*	*F-max*	*F-max*	*AUROC*	*AUROC*	*Sp.*	*Sp.*	*Sp.*	
ProteinMPNN	0.0116	0.0056	0.0039	0.0020	0.0005	0.0031	0.0036	0.0080	0.0493	0.1254	26.7
Foldseek	0.0209	0.0036	0.0022	0.0038	0.0014	0.0040	0.0036	0.0040	0.0530	0.0127	26.9
TFN	0.0400	0.0053	0.0098	0.0035	0.0021	0.0054	0.0040	0.0042	0.0539	0.0250	27.3
GearNet-pre	0.0089	0.0076	0.0045	0.0004	0.0015	0.0330	0.0024	0.0004	0.0068	0.0731	29.1
IPA-pre	0.0130	0.0061	0.0064	0.0025	0.0016	0.0062	0.0048	0.0067	0.0227	0.0271	30.9
composition reference	0.0046	0.0042	0.0072	0.0036	0.0026	0.0091	0.0004	0.0049	0.0508	0.2862	–

K Latent-Space Atlas

Does the downstream ordering of Section 3.2 appear without a trained head? The atlas answers this question. It maps where every encoder places the same proteins, with one layer per scored downstream task, so that each map can be read next to the score the same encoder obtains on that task. Section 3.4 cites it for one claim: on every task layer, the encoder whose nearest neighbours most often share a protein's label reads sequence only. The atlas is otherwise descriptive. It shows two further results: encoders that score similarly on probes can organise the same proteins differently, and a representation can look healthy in aggregate while being degenerate. The atlas data (`atlas/data_v4/atlas.json`) and the scripts that build and draw it are released with the benchmark.

Construction. Every layer is built from the released embeddings, extracted from each encoder's audited inputs (`protcompass/extract/embeddings`) and mean-pooled per protein. The rows are those of the train and test splits the task was scored on, so every row is a vector a downstream head was trained or tested on. Every layer covers the 33 encoders of the paper (32 on remote homology, where FoldVision has no extraction). Each layer is a stratified sample of about 2,000 proteins (seed 42) that takes at least 150 proteins from each class, or the whole class where it is smaller, and each protein is labelled by the task's own target. *EC classification*: 2,242 ECReaction chains, one per distinct sequence, labelled by top-level EC class (EC 1 307, EC 2 642, EC 3 668, EC 4 186, EC 5 150, EC 6 150, EC 7 139), with class names recovered by a Swiss-Prot vote over each ProteinWorkshop class index. *GO-MF* and *GO-BP*: the annotations are multi-label, with 6.3 and 31.2 terms per chain on average and the frequent terms nested in one another, so we chose eight frequent terms per branch, none nested in another (MF: ATP binding, catalytic activity acting on a protein, transition metal ion binding, RNA binding, DNA binding, structural molecule activity, transporter activity, peptidase activity; BP: transport, catabolic process, RNA metabolic process, response to stress, cell communication, macromolecule modification, organic acid metabolic process, cellular component organization) and label each chain by the least frequent of these it carries, or "other"; 2,215 MF chains (other 985) and 2,087 BP chains (other 609). *Localisation*: 2,362 DeepLoc 2 proteins labelled by compartment (Cytoplasm 689, Nucleus 405, Extracellular 224, Cell membrane 204, Mitochondrion 169, four compartments at 150 and all 71 Peroxisome proteins). *Remote homology*: 2,129 SCOP domains labelled by SCOP class (all alpha 369, all beta 487, alpha/beta 455, alpha+beta 474, multi-domain 98, membrane and cell surface 92, small proteins 154). *Solubility*: 1,999 proteins, 1,389 soluble and 610 insoluble. *Membrane versus soluble*: 1,999 proteins, 547 membrane and 1,452 soluble. *Stability* and *fluorescence*: 1,998 and 2,000 proteins labelled by quartile of the target over the sample. *PPI affinity* is a property of a chain pair, so its layer holds 1,998 chains, each labelled by the quartile of the mean affinity of the pairs it takes part in (2,872 pairs). Each encoder's embedding is reduced by UMAP with $k = 10$ neighbours and seed 42, then Procrustes-aligned to a common reference (ESM2-150M); the stored coordinates are centred and scaled to unit RMS radius, and every panel of a layer uses the same frame (Figures 11, 12, 13, 14, 15, 16, 17, 18, 19 and 20). Similarity between encoders is measured by debiased linear CKA and by the overlap of k-nearest-neighbour sets, both computed on the full embeddings, so no conclusion depends on the layout algorithm or the 2D projection. Colours follow the main figures: panel borders and encoder names are navy for sequence-based, orange for structure-based and purple for multimodal encoders, using the declared classes of Section 2; the CKA matrices use the colour map of Figure 2, and protein labels use a muted categorical set. The layers, figures and summary table are produced by `protcompass/scripts/build_atlas_v4.py`, `make_atlas_figures_paper.py` and `emit_atlas_summary.py`.

Table 48: Latent-space atlas, one layer per scored downstream task. Each layer is a stratified sample of the proteins the task was scored on, labelled by the task's own target (Appendix K). CKA is debiased linear CKA between encoders on the full embeddings, averaged over the off-diagonal pairs; within/between is by declared class; rand/pre are the randomly initialised and pretrained structure encoders. k-NN purity ($k = 10$) and the five-fold linear probe are given as the range over the encoders with the encoder at each end, against chance and the majority-class rate.

layer	n	classes	CKA mean (range)	within / between	rand–rand / pre–pre / rand–pre
EC classification	2,242	7	0.290 (0.02–0.93)	0.336 / 0.259	0.535 / 0.281 / 0.214
GO-MF	2,215	9	0.288 (0.03–0.95)	0.339 / 0.252	0.533 / 0.310 / 0.227
GO-BP	2,087	9	0.287 (0.02–0.94)	0.333 / 0.254	0.527 / 0.303 / 0.219
Localisation	2,362	10	0.397 (0.03–0.94)	0.442 / 0.365	0.574 / 0.467 / 0.323
Remote homology	2,129	7	0.321 (0.02–0.93)	0.353 / 0.297	0.553 / 0.270 / 0.219
Solubility	1,999	2	0.360 (0.03–0.94)	0.393 / 0.336	0.540 / 0.399 / 0.254
Membrane vs soluble	1,999	2	0.396 (0.03–0.94)	0.441 / 0.366	0.569 / 0.466 / 0.322
Stability	1,998	4	0.484 (0.06–0.95)	0.474 / 0.492	0.609 / 0.404 / 0.297
Fluorescence	2,000	4	0.207 (0.00–0.91)	0.282 / 0.155	0.345 / 0.352 / 0.243
PPI affinity	1,998	4	0.618 (0.16–0.94)	0.592 / 0.636	0.752 / 0.432 / 0.483

layer	purity min	purity max	chance	probe min–max (majority)
EC classification	0.289 (IPA-pre)	0.802 (Ankh)	0.209	0.362–0.937 (0.298)
GO-MF	0.259 (IPA-pre)	0.559 (Ankh)	0.236	0.333–0.708 (0.445)
GO-BP	0.168 (GearNet-pre)	0.367 (Ankh)	0.151	0.194–0.486 (0.292)
Localisation	0.179 (IPA-pre)	0.467 (ProtT5-XL)	0.153	0.275–0.624 (0.292)
Remote homology	0.282 (IPA-pre)	0.748 (Ankh)	0.187	0.402–0.925 (0.229)
Solubility	0.571 (Foldseek)	0.628 (ProtT5-XL)	0.576	0.598–0.696 (0.695)
Membrane vs soluble	0.635 (IPA-pre)	0.762 (Ankh)	0.603	0.682–0.825 (0.726)
Stability	0.394 (IPA-pre)	0.485 (ESM2-150M)	0.250	0.449–0.537 (0.254)
Fluorescence	0.276 (IPA-pre)	0.346 (ProstT5)	0.250	0.316–0.460 (0.250)
PPI affinity	0.482 (IPA-pre)	0.596 (ESM2-650M)	0.250	0.513–0.672 (0.250)

Class structure tracks the downstream score. The k-NN purity of a panel is the fraction of each protein's ten nearest neighbours that share its label, computed on the full embedding. It reads the label structure of a representation without a trained head. The encoder with the highest purity is a sequence-based model on nine layers (Ankh on EC, GO-MF, GO-BP, remote homology and membrane; ProtTrans-T5 on localisation and solubility; ESM2-150M on stability; ESM2-650M on affinity) and ProstT5 on fluorescence, which is multimodal by declared class and reads sequence only by the input audit, so by the audit the encoder with the highest purity reads sequence only on all ten layers. On every layer the five highest purities belong to sequence-only encoders, ESM3, ProstT5 or SaProt, with ESM-IF1 in fifth place on remote homology and fluorescence. The lowest purity is IPA-pretrained on eight layers, GearNet-pretrained on GO-BP and Foldseek 3Di on solubility, where it sits at chance (0.571 against 0.576); every other encoder on every layer is above chance. The largest spreads are on the tasks the sequence-only group leads by most in Section 3.2: EC (0.289 to 0.802 against chance 0.209), remote homology (0.282 to 0.748 against 0.187) and localisation (0.179 to 0.467 against 0.153); on the two binary layers and the three regression layers the spread is 0.06 to 0.13. A five-fold linear probe on the same rows follows the same ordering (Table 48); on the regression layers a five-fold ridge readout of the target reaches Spearman 0.54 to 0.74 on stability, 0.29 to 0.72 on fluorescence and 0.47 to 0.85 on the chain-mean affinity.

Encoders that probe alike can organise proteins differently. Debiased linear CKA between two encoders measures how alike they organise the same proteins (Table 48). The mean off-diagonal CKA between encoders ranges from 0.207 on the fluorescence layer to 0.618 on the affinity layer, and on every layer the range across encoder pairs runs from near 0 to above 0.9. Two encoders can therefore assign the same proteins to different neighbourhoods while scoring alike on every probe. The declared class explains some of this structure: CKA within a class exceeds CKA between classes

on eight of the ten layers, by 0.05 to 0.13; the exceptions are the two regression layers on nearly identical sequences, stability (0.474 against 0.492) and affinity (0.592 against 0.636). Training status explains more. Among the structure encoders, randomly initialised networks resemble one another (CKA 0.53 to 0.75 on nine layers, 0.35 on fluorescence) more than pretrained networks resemble one another (0.27 to 0.47) and more than a random network resembles a pretrained one (0.21 to 0.48); the five architectures present in both states split the same way on every layer, GearNet and IPA sharing a CKA of 0.08 to 0.26 between their random and pretrained versions (IPA 0.40 on affinity) and PottsMPNN, GCPNet and TFN 0.27 to 0.82. The randomly initialised geometric networks form one block of the CKA matrix on every layer (Figure 8; Figures 9 and 10 show two layers with the encoders named).

The maps expose degenerate embeddings. The maps also show when an extraction has failed. In a set of cached embeddings that the paper does not use, they made the failures of four encoders obvious, and each failure has a diagnosis: dMaSIF returned all-zero embeddings; a fine-tuned PottsMPNN checkpoint collapsed to two distinct vectors across 5,000 sampled proteins; the cached IPA entry did not cover the sampled set; and the GCPNet wrapper built a new unseeded random network for every protein, so its rows were not mutually comparable and its k-NN purity sat at 0.28 against a chance level of 0.25. Every encoder in the release is extracted from audited inputs. In the released atlas every panel has as many distinct rows as proteins, up to duplicate sequences, with one explained exception: the 20-dimensional Foldseek 3Di state composition takes the same value on many short chains of the affinity layer (1,090 distinct rows of 1,998) and on point variants of one design in the stability layer (1,661 of 1,998). Reading k-NN purity against chance catches all four failures, and we recommend it as a routine sanity check on any released embedding set.

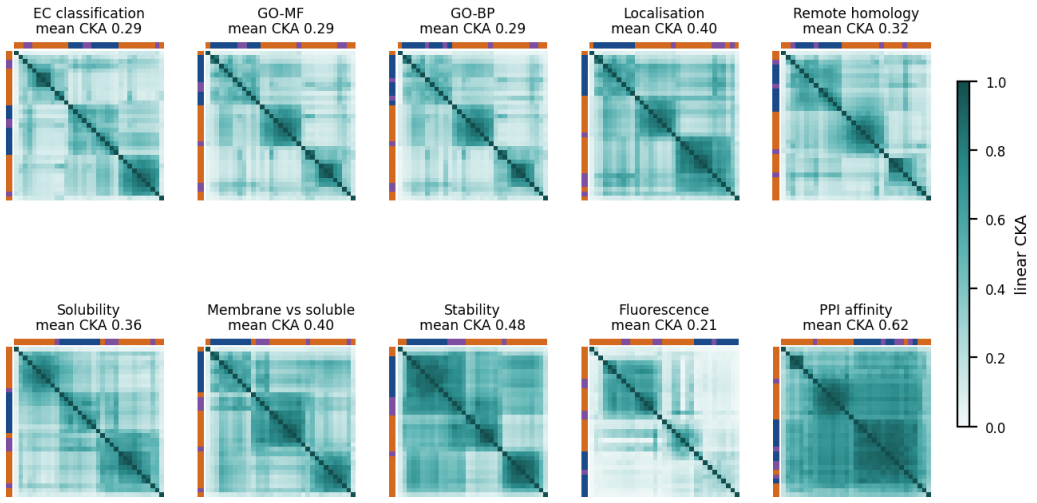
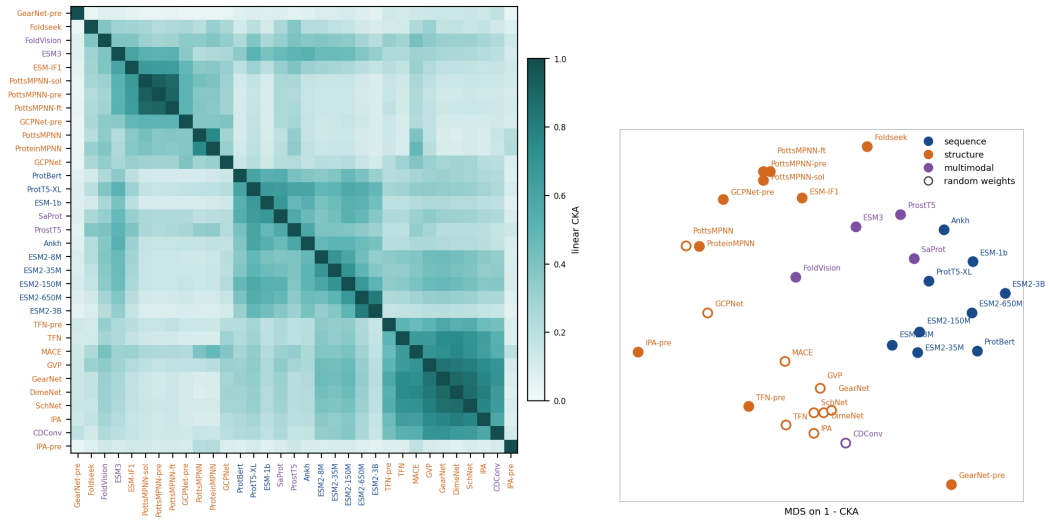
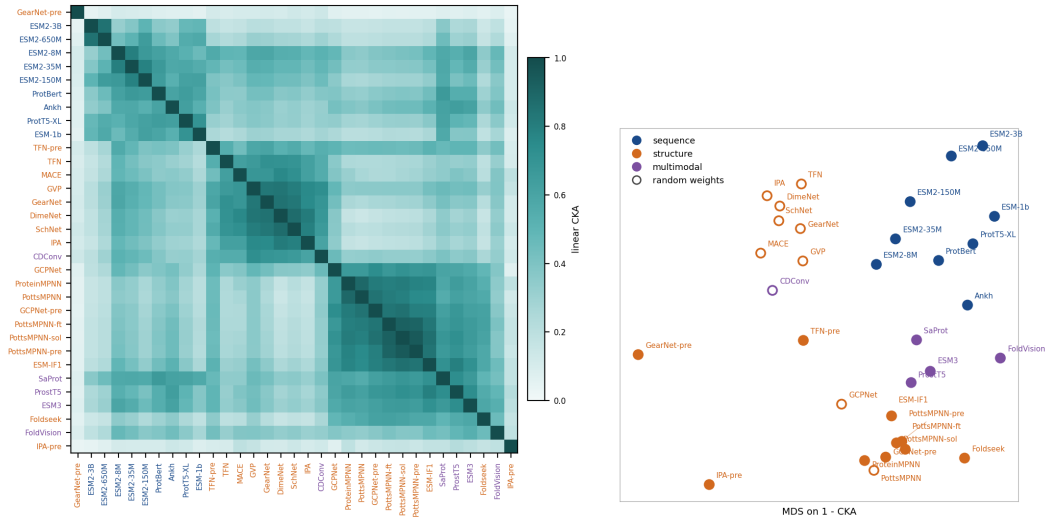


Figure 8: Encoder similarity on every layer. Debiased linear CKA between the 33 encoders (32 on remote homology) on the full embeddings of each layer, each matrix ordered by its own average-linkage clustering; the strips along the edges give the declared class of each encoder (navy sequence-based, orange structure-based, purple multimodal). The block of randomly initialised geometric networks appears on every layer.



(a) Debiased linear CKA, EC classification layer **(b)** Encoders embedded by $1 - \text{CKA}$

Figure 9: Encoder similarity on the EC classification layer, with the encoders named. CKA is computed on the full embeddings and ordered by average-linkage clustering; tick colours give the encoder class. Open circles are randomly initialised networks, which cluster with each other and away from their pretrained counterparts.



(a) Debiased linear CKA, localisation layer **(b)** Encoders embedded by $1 - \text{CKA}$

Figure 10: Encoder similarity on the localisation layer, with the encoders named. Same conventions as Figure 9.



Figure 11: Latent-space atlas, EC classification layer. UMAP maps of the same 2,242 ECReaction chains under all 33 released encoders, coloured by top-level EC class. Panel border and name colour give the encoder class (navy sequence-based, orange structure-based, purple multimodal); "rand." marks randomly initialised weights; within each class the panels are ordered by k-NN purity. Every panel is Procrustes-aligned to ESM2-150M and drawn in one common frame.

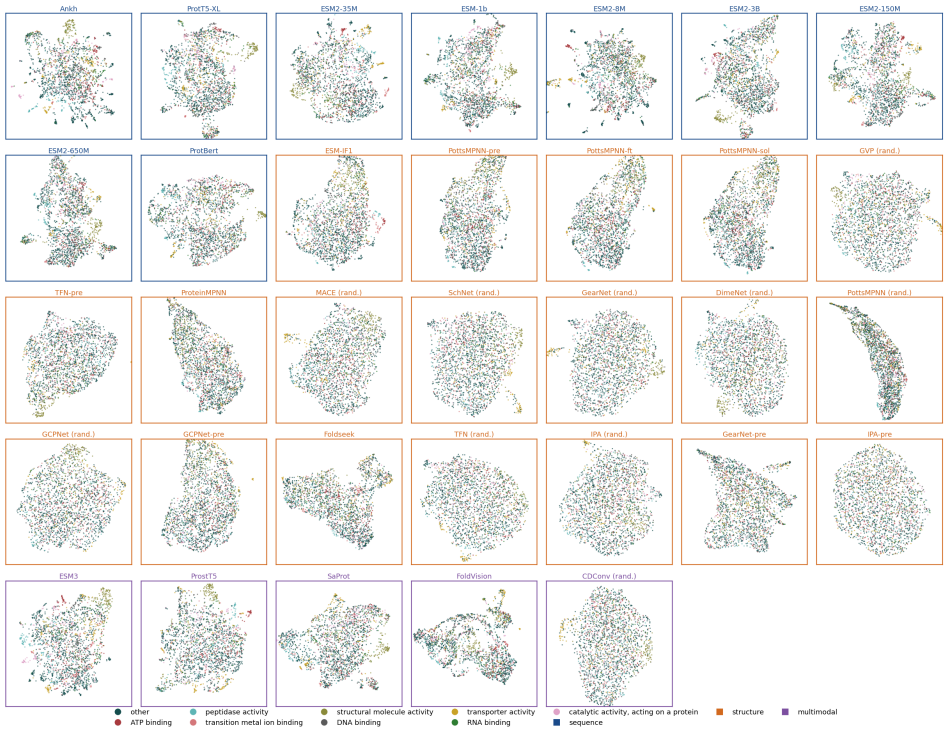


Figure 12: Latent-space atlas, GO molecular function layer. 2,215 nrPDB-GO chains coloured by the least frequent of eight frequent MF terms they carry ("other" if none). Same conventions as Figure 11.



Figure 13: Latent-space atlas, GO biological process layer. 2,087 nrPDB-GO chains coloured by the least frequent of eight frequent BP terms they carry ("other" if none). Same conventions as Figure 11.



Figure 14: Latent-space atlas, localisation layer. 2,362 proteins of the DeepLoc 2 task coloured by compartment. Same conventions as Figure 11.

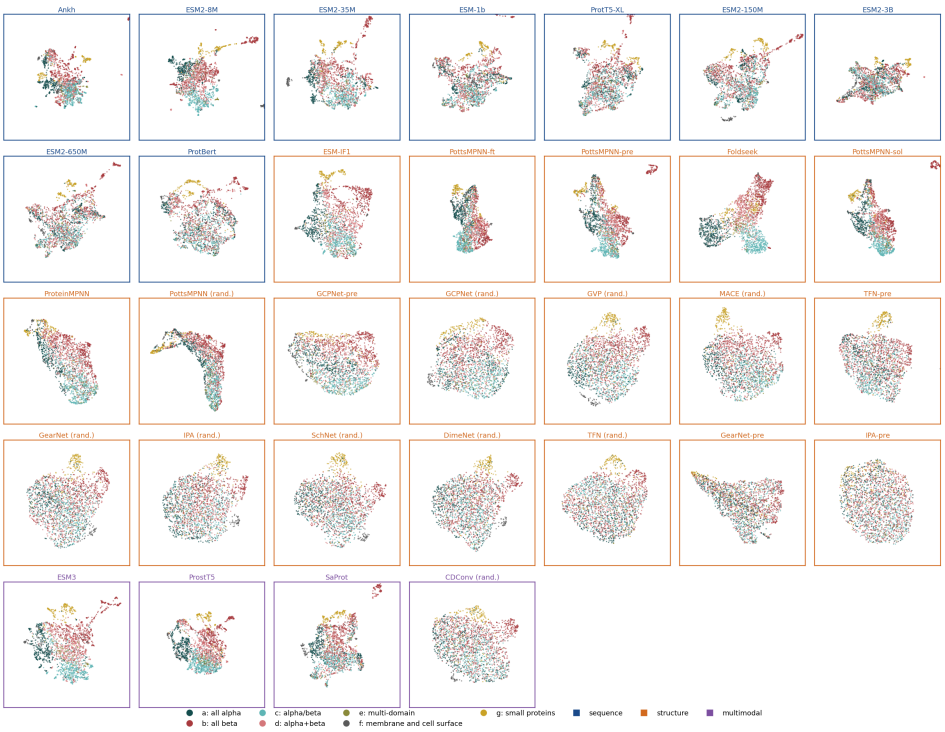


Figure 15: Latent-space atlas, remote homology layer. 2,129 SCOP domains of the TAPE task coloured by SCOP class; 32 encoders, because FoldVision has no extraction on this task. Same conventions as Figure 11.

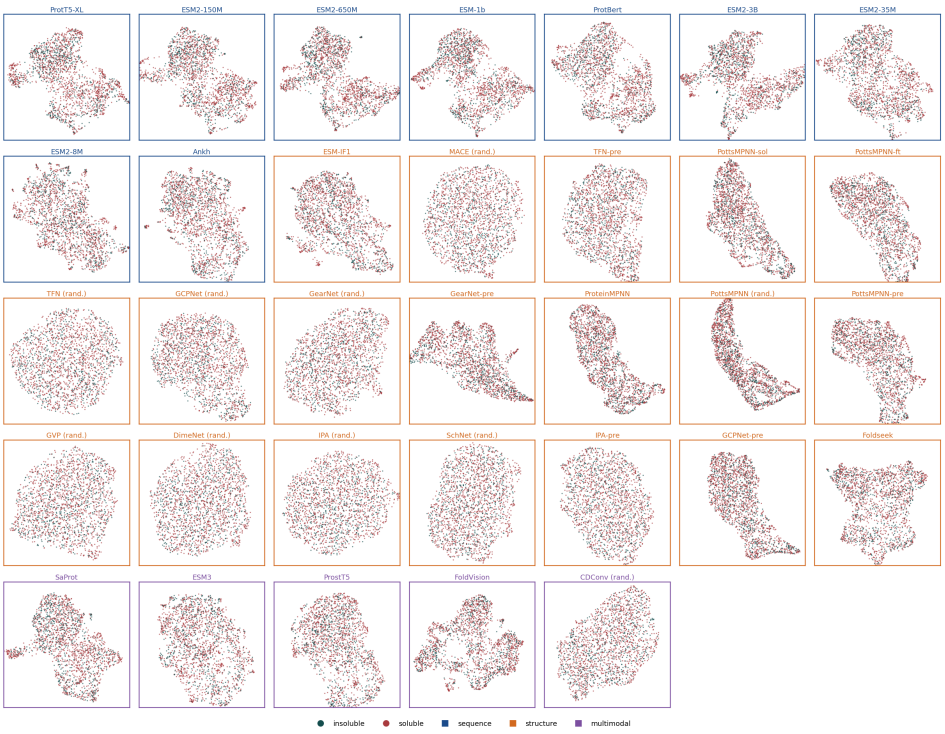


Figure 16: Latent-space atlas, solubility layer. 1,999 proteins of the DeepSol task coloured soluble or insoluble. Same conventions as Figure 11.



Figure 17: Latent-space atlas, membrane versus soluble layer. 1,999 proteins of the DeepLoc 2 membrane task coloured membrane or soluble. Same conventions as Figure 11.

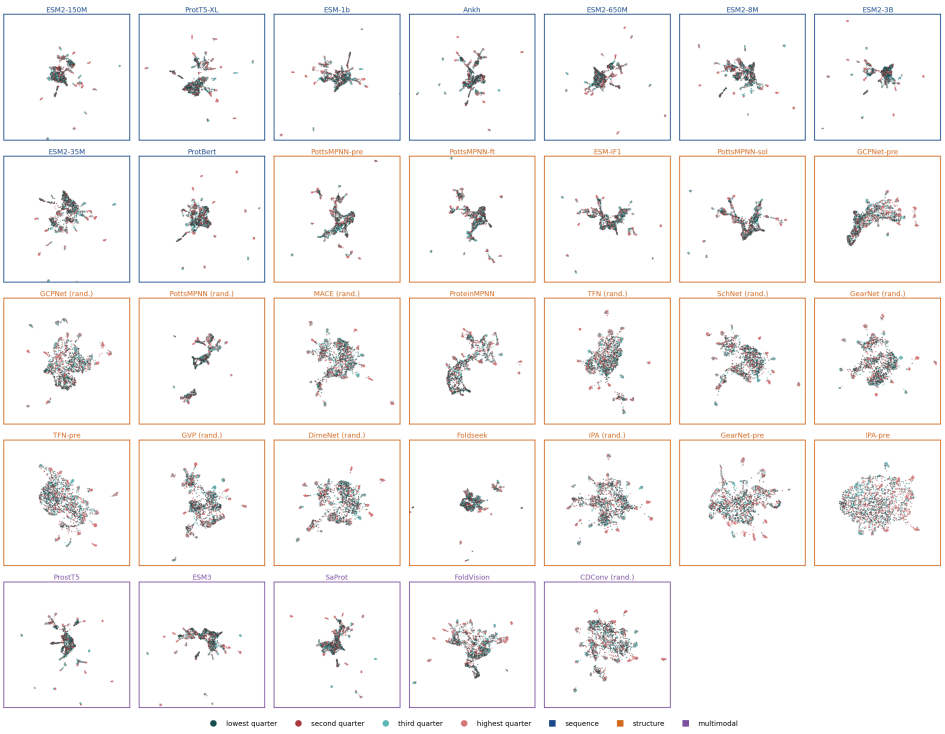


Figure 18: Latent-space atlas, stability layer. 1,998 designed mini-proteins and variants coloured by quartile of the stability score. Same conventions as Figure 11.

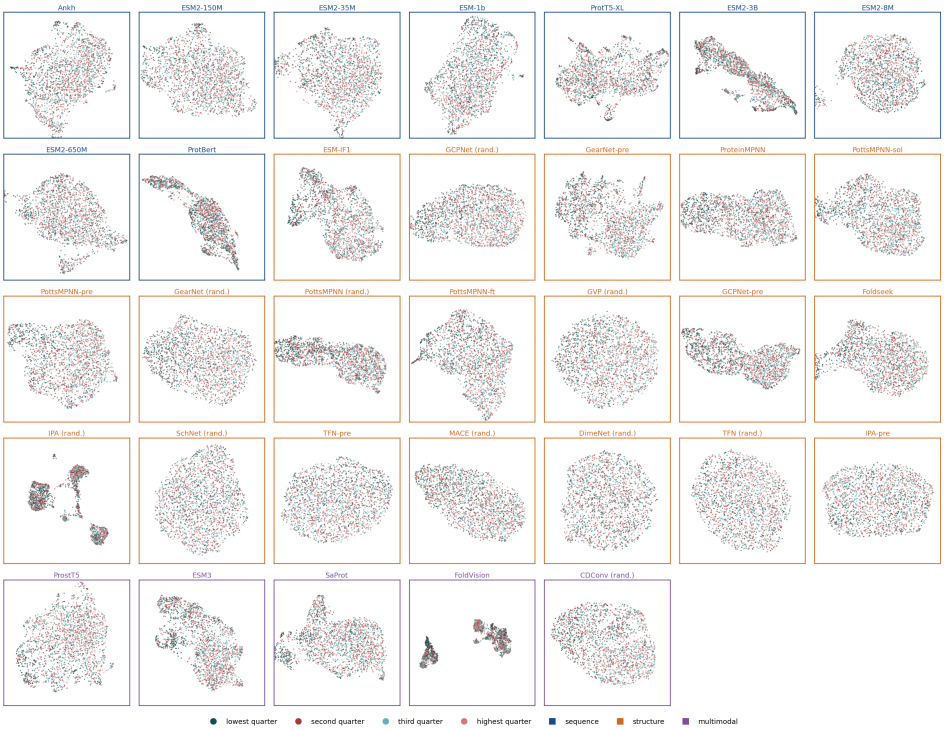


Figure 19: Latent-space atlas, fluorescence layer. 2,000 GFP variants coloured by quartile of log fluorescence. Same conventions as Figure 11.

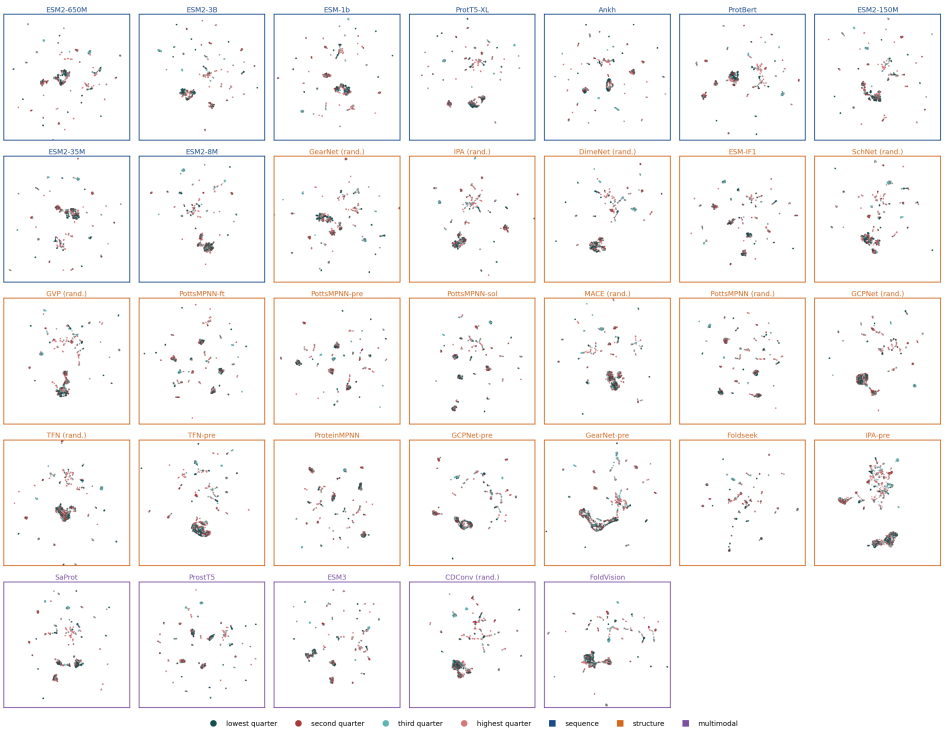


Figure 20: Latent-space atlas, PPI affinity layer. 1,998 chains of the SKEMPI-derived affinity task coloured by quartile of the mean affinity of the pairs each chain takes part in. Same conventions as Figure 11.