OMP: ORDER-INVARIANT PROTEIN INTERACTION LANGUAGE MODELING WITH LAYER-WISE SIAMESE CROSS-ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein language models (PLMs) adapted to protein–protein interactions (PPIs) typically concatenate the two protein input sequences, which forces aggressive cropping, makes predictions depend on input order, and requires every residue to split a single attention budget between its own chain and its partner. We introduce OMP, a dual encoder that extends ESM-2 (650M) with Siamese cross-attention modules running in parallel to self-attention in every layer, each followed by a gated linear unit that is initialized nearly closed. OMP processes up to 1,022 residues per chain, twice the per-chain context of concatenation-based models built on the same backbone, and is order-invariant by construction. We pre-train OMP with masked language modeling on interacting sequence pairs from STRING and evaluate its frozen embeddings on three downstream tasks. Pre-training takes 4,032 GPU-hours over a fifth of the curated corpus, so adding pair-level structure to a pre-trained PLM does not require re-pretraining at the scale of the backbone. On the gold-standard PPI benchmark, OMP reaches 66.5% accuracy and 0.712 AUPRC, surpassing the 65% accuracy plateau of sequence-based models on this leakage-controlled benchmark and significantly outperforming ESM-2 with an identical prediction head. Among the models evaluated here, only OMP exceeds 0.60 mean AUROC on TCR–epitope binding prediction, and it outperforms the interaction language model MINT on interface prediction, where the single-sequence ESM-2 baseline remains strongest. Without structural supervision, OMP's attention maps recover protein domains: 86% of the strongest intra-protein energetic contacts in a target protein fall within the top 15% of layer-30 self-attention. Anonymized code and pre-trained weights are available at https://anonymous.4open.science/r/OMP and https://huggingface.co/OMP-ICLR/OMP.

1 INTRODUCTION

Proteins are the primary macromolecules driving cellular processes, and determining whether two amino acid sequences bind provides profound insights for biological research. Protein language models (PLMs) and structure predictors such as the AlphaFold family (Abramson et al., 2024) and the open-source Boltz-2 (Passaro et al., 2025) process sequences with attention mechanisms analogous to those used for natural language (Vaswani et al., 2017), predict the structures of millions of proteins, and can guide the design of sequences that promote or inhibit target interactions (Hayes et al., 2025). However, adapting them to protein–protein interactions (PPIs) remains computationally costly and typically relies on concatenating the two sequences, with the drawbacks we now explain.

A natural route toward a protein interaction language model is to pre-train on pairs of interacting sequences. Current interaction language models such as MINT (Ullanat et al., 2026), PPLM (Liu et al., 2026), and PLM-interact (Liu et al., 2025) concatenate the two sequences into a single input, an architectural choice with three major problems. First, both chains share one context window, so long proteins must be cropped to a point where models miss a substantial amount of data: MINT truncates each sequence at 512 residues, which alters 44.32% of the 11,018 unique chains of the gold-standard PPI dataset (Bernett, 2022), and long chains are common enough that 14.45% of them

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.

exceed even the 1,022 residues that OMP admits per chain. Second, predictions depend on the order of the chains: with the released MINT checkpoint and its gold-standard prediction head, swapping the two inputs flips the predicted label for 16 of the 30 pairs we tested (Appendix E). Third, every residue must split a single softmax attention budget between its own chain and its partner.

We introduce OMP, a dual encoder that extends ESM-2 (650M) (Lin et al., 2023) with a Siamese cross-attention module running in parallel to self-attention in every layer, so that each protein enriches its own representation by querying its partner. OMP does not eliminate cropping, but it mitigates it: each chain receives the full 1,022-residue context window of the backbone rather than sharing one with its partner, which doubles the per-chain budget of concatenation-based models built on the same backbone and leaves only the 14.45% of chains longer than that to be cropped. It also keeps intra- and inter-chain attention budgets independent and is order-invariant by construction. Although similar cross-attention designs have been used in downstream classification heads (Agassy et al., 2025; Reim et al., 2025), our results indicate that integrating this scheme across every layer of the base transformer during pre-training improves performance across diverse downstream tasks. We evaluate OMP with frozen embeddings on three tasks in which proteins longer than 1,000 amino acids are common, and summarize our contributions as follows:

- **PPI prediction.** On the gold-standard PPI dataset (Bernett, 2022), OMP reaches 66.5% accuracy and 0.712 AUPRC. To our knowledge, it is the first sequence-based model with a reported accuracy above the 65% plateau on this benchmark. The plateau is specific to benchmarks with strict sequence deduplication and degree-balanced negatives: on datasets split at random, sequence redundancy and node degree inflate the reported performance, and models that appear highly accurate there fall to near-random once those shortcuts are removed (Bernett et al., 2024). The AUPRC gain over ESM-2 under an identical prediction head is significant (99% CI [0.0065, 0.0150]).

- **Domain-specific and residue-level tasks.** Among the general-purpose models we evaluate and the TDC-2 baselines, OMP is the only one with a mean AUROC above 0.60 on TCR–epitope binding prediction from TDC-2 (Velez-Arce et al., 2024); specialized TCR models, which are trained end-to-end for the domain and report results under different setups, are discussed in Appendix H. OMP also outperforms MINT on interface prediction (Wilcoxon $p = 0.0086$).

- **Cost.** Warm-started from ESM-2, OMP is pre-trained in 4,032 GPU-hours over 19% of the curated STRING corpus, so equipping a pre-trained PLM with pair-level structure does not require re-pretraining at the scale of the backbone.

- **Interpretability.** Without structural supervision, the attention maps of OMP recover protein domains, while preserving the correspondence between self-attention and BLOSUM62 substitution scores inherited from ESM-2 (Vig et al., 2021).

2 Related Work

2.1 Protein language models for PPI prediction

To avoid the cost of multiple sequence alignments, many PPI predictors rely on embeddings from lighter single-sequence encoders such as ESM-2 (Lin et al., 2023) and ProtT5 (Elnaggar et al., 2022), with classification heads over pooled protein descriptors (Sledzieski et al., 2024). Structure-aware encoders such as SaProt (Su et al., 2024) integrate structural tokens to improve pre-training perplexity and downstream PPI accuracy. Because global mean or max pooling discards the resolution needed for interaction prediction, D-SCRIPT (Sledzieski et al., 2021), TUnA (Ko et al., 2024), Topsy-Turvy (Singh et al., 2022), and xCAPT5 (Dang & Vu, 2024) operate on full per-residue embeddings combined through attention or convolution, but still depend on cropping or padding. Single-sequence encoders appear insufficient for robust PPI prediction (Reim et al., 2025), even when supplemented with functional annotations, text, or structural data (Wu et al., 2024; Su et al., 2024), which motivates learning directly from interacting pairs.

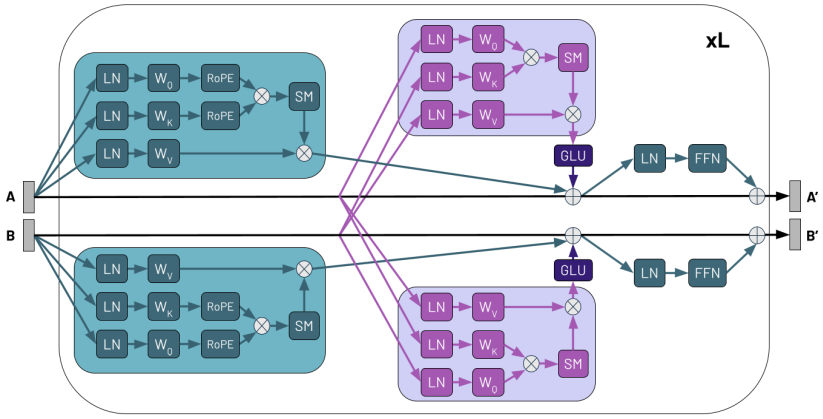


Figure 1: Architecture of OMP. The Siamese parallel cross-attention modules integrated in every layer of the encoder are highlighted in pink. Rotary positional embeddings (RoPE) are applied only in the self-attention blocks (teal), with positions counted within each chain; the cross-attention blocks use no positional encoding.

2.2 Interaction language models

MINT (Ullanat et al., 2026) re-trains ESM-2 on interacting pairs from the STRING database (Szklarczyk et al., 2023): the two sequences are concatenated, and layer-wise masking lets one attention module capture inter-protein features while a parallel module tracks intra-protein properties. PPLM (Liu et al., 2026) and PLM-interact (Liu et al., 2025) also concatenate the inputs and explore strategies to mark chain boundaries, and PLM-interact augments training data by repeating pairs in the opposite order. OMP instead keeps the two chains in separate streams, so its behavior depends neither on boundary tokens nor on order augmentation. ProteomeLM (Malbranke et al., 2026) changes the unit of modeling instead, operating on whole proteomes and recovering interactions from its attention coefficients without interaction labels. Cross-attention between partner embeddings has been used in PPI classification heads on frozen encoders (Agassy et al., 2025; Reim et al., 2025), and in task-specific predictors for our downstream domains: ProtLM.TCR (Essaghir et al., 2022) stacks self- and cross-attention over separately embedded TCR and epitope sequences, LANTERN (Qi et al., 2026) couples an ESM encoder for the TCR with a molecular encoder for the peptide, and Pair-EGRET (Alam et al., 2024) applies it over a Siamese network on PLM features and a structural graph. All place the operation in a predictor trained for one domain, whereas OMP learns it inside the encoder during pre-training and is then evaluated frozen across three tasks (Appendix H).

2.3 Interpretability of protein language models

Vig et al. (2021) showed that attention heads in PLMs specialize on amino acid types and that attention similarity between residues correlates with BLOSUM62 substitution scores (Henikoff & Henikoff, 1992). Raw attention and contact maps are nonetheless difficult to interpret directly (Reim et al., 2025), and structural models rely on Predicted Aligned Error (PAE) matrices (Abramson et al., 2024; Passaro et al., 2025), an indirect proxy that requires supervised structural training rather than emerging from sequence pre-training. In large transformers, attention sinks (Xiao et al., 2024) and massive activations (Sun et al., 2024) can further distort attention-based interpretation, and gated attention has been shown to mitigate these artifacts (Qiu et al., 2025).

3 Method

3.1 Siamese parallel cross-attention

Let $\boldsymbol{H}_a \in \mathbb{R}^{n_a \times d}$ and $\boldsymbol{H}_b \in \mathbb{R}^{n_b \times d}$ denote the token representations of two interacting protein chains of lengths n_a and n_b. A model restricted to the concatenation $[\boldsymbol{H}_a; \boldsymbol{H}_b]$ processes interactions through standard self-attention (projections, multiple heads, and rotary embeddings are omitted for

clarity; rotary embeddings are discussed in Sections 3.2 and 3.5):

$$\text{Attention}([\boldsymbol{H}_a; \boldsymbol{H}_b]) = \text{softmax}\left(\frac{[\boldsymbol{H}_a; \boldsymbol{H}_b][\boldsymbol{H}_a; \boldsymbol{H}_b]^T}{\sqrt{d_k}}\right)[\boldsymbol{H}_a; \boldsymbol{H}_b], \tag{1}$$

where the softmax is applied row-wise and d_k is the per-head key dimension of the omitted projections. In block form, the attention scores are

$$\boldsymbol{S} = \frac{1}{\sqrt{d_k}}\begin{pmatrix} \boldsymbol{H}_a\boldsymbol{H}_a^T & \boldsymbol{H}_a\boldsymbol{H}_b^T \\ \boldsymbol{H}_b\boldsymbol{H}_a^T & \boldsymbol{H}_b\boldsymbol{H}_b^T \end{pmatrix}, \tag{2}$$

so intra-chain scores lie on the block diagonal and inter-chain scores on the off-diagonal blocks. Because the softmax normalizes each row of $\boldsymbol{S}$ jointly over both blocks, each token must split a single attention budget between its own chain and its partner. OMP disentangles these operations by adding, in every layer, a cross-attention module in parallel to the original self-attention encoder, with its own query, key, value and output projections and its own layer normalization, sharing no weights with self-attention and using the same number of heads (Figure 1):

$$\begin{aligned} \text{Attention}(\boldsymbol{H}_a \to \boldsymbol{H}_b) &= \text{softmax}\left(\frac{\boldsymbol{H}_a\boldsymbol{H}_b^T}{\sqrt{d_k}}\right)\boldsymbol{H}_b, \\ \text{Attention}(\boldsymbol{H}_b \to \boldsymbol{H}_a) &= \text{softmax}\left(\frac{\boldsymbol{H}_b\boldsymbol{H}_a^T}{\sqrt{d_k}}\right)\boldsymbol{H}_a, \end{aligned} \tag{3}$$

where each softmax is normalized over the residues of the partner chain only. The same (Siamese) modules process both chains, which solves the three problems of concatenation. First, each chain can use the full context window of the backbone, twice the per-chain context of current concatenation-based models (Ullanat et al., 2026), while the number of attention scores, $n_a^2 + n_b^2 + 2n_an_b = (n_a + n_b)^2$, remains that of attention over the concatenated pair. The score matrices therefore cost the same, but the projections do not: every layer computes a second set of queries, keys and values for the cross-attention stream, which doubles the projection cost of the attention blocks and, together with the gate of the GLU, raises the matrix-multiplication cost of a layer by roughly half at equal total length, in the forward and in the backward pass (Appendix A). Second, the architecture guarantees order invariance (Section 3.5) and is therefore insensitive to the arbitrary order conventions of asymmetric datasets. Finally, each token receives independent attention budgets for intra- and inter-chain information.

3.2 Positional encoding and gating

OMP handles positions differently in its two attention modules. The self-attention modules keep the rotary positional embeddings (RoPE) of ESM-2 (Lin et al., 2023), with positions restarting at the beginning of each chain, so the positions of a chain do not depend on the order in which the pair is given. The cross-attention modules use no positional encoding, reflecting that interacting macromolecules bind without a strict dependence on the relative positions of residues across different chains; positional information reaches cross-attention only through the hidden states produced by self-attention. Finally, cross-attention outputs are processed through gated linear units (GLUs) (Agassy et al., 2025; Qiu et al., 2025), before entering the residual stream alongside the self-attention output. For the cross-attention output $\boldsymbol{c}_i \in \mathbb{R}^d$ of residue i,

$$[\boldsymbol{z}_i; \boldsymbol{g}_i] = \boldsymbol{W}\boldsymbol{c}_i + \boldsymbol{b}, \qquad \text{GLU}(\boldsymbol{c}_i) = \boldsymbol{z}_i \odot \sigma(\boldsymbol{g}_i), \tag{4}$$

where $\boldsymbol{W} \in \mathbb{R}^{2d\times d}$, σ is the logistic sigmoid, and $\sigma(\boldsymbol{g}_i)$ is the gate openness. The gate half of the bias $\boldsymbol{b}$ is initialized to -6, so the GLUs start nearly closed ($\sigma(-6) \approx 0.0025$) and gradually admit cross-attention features into the residual stream as training matures, which stabilizes pre-training and transfer learning from the ESM-2 weights.

The backbone is pre-LayerNorm, and both branches read the same layer input: self-attention through its own normalization and cross-attention through its own, whose output passes through the GLU, and the two are summed into the residual stream in a single step before the pre-norm feed-forward block. The cross-attention branch therefore does not reuse the normalization of the self-attention branch, its queries come from one chain and its keys and values from the partner, and every module is shared between the two chains; Eq. (5) in Appendix A writes the update out.

3.3 PRE-TRAINING

We warm-start the self-attention and feed-forward modules from ESM-2 (650M) (Lin et al., 2023), for a total of 975M trainable parameters, and pre-train OMP with masked language modeling on interacting pairs from STRING (Szklarczyk et al., 2023). Following a curation protocol similar to MINT (Ullanat et al., 2026), we clustered sequences at 50% identity and removed validation families from the training set, obtaining 96,420,565 training interactions across 16,392,879 unique sequences (Appendix B.1). Due to computational constraints, training ran in one-week intervals that resumed from the latest checkpoint but restarted data sampling from the beginning of the dataset, rather than resuming the state of the sampler. The model therefore saw the same subset of 18,421,200 pairs five times, that is 19% of the curated interactions, instead of a larger sample once; the best checkpoint was obtained at the end of the fourth epoch. Validation perplexity and GLU openness were tracked throughout pre-training (Figure 4; Appendix A.1).

3.4 DOWNSTREAM PREDICTION HEADS

In all downstream tasks the pre-trained backbone remains frozen, so only a lightweight head is trained. For pair-level classification, an affinity module appends one trainable Siamese block and recombines the chains in an order-invariant manner (Figure 5). The block is not a full OMP layer: it carries cross-attention with its own layer normalization and a pre-norm feed-forward network, but neither self-attention nor a gated linear unit, and Eq. (6) of Appendix A.2 writes it out, with all parameters shared by the two chains and no positional encoding. Following established strategies (Reim et al., 2025; Passaro et al., 2025), the post-cross-attention embeddings of each protein are pooled by mean and maximum over its residue positions, excluding padding and special tokens, into a 2,560-dimensional representation; the two are recombined through element-wise Hadamard products and absolute differences, and a shallow multi-layer perceptron (MLP) outputs the interaction probability. For residue-level interface prediction, an MLP is applied to the final embedding of each amino acid. Hyperparameters are listed in Appendix A.2.

3.5 ORDER INVARIANCE

OMP is order-invariant by construction: its predictions do not depend on which protein is given first. The argument rests on five structural properties of the encoder, verifiable in the released implementation and listed in Appendix G: separate embedding with no chain-identity token, RoPE only in self-attention with per-chain positions, parameters shared by both chains, padding masked and pooling restricted to residue positions, and simultaneous updates from the partner's previous-layer state. Under them every layer applies one shared function to a chain and its partner, so the encoder is equivariant under swapping the pair, and the affinity module and pooling preserve that equivariance while the symmetric recombination turns it into invariance. Write $\Phi(a,b) = (\boldsymbol{H}_a^{(L)}, \boldsymbol{H}_b^{(L)})$ for the representations the encoder produces after its L layers and Φ_{aff} for the Siamese block of the affinity module (Section 3.4).

Theorem 1 (Order invariance of OMP)**.** *Let* $(\tilde{\boldsymbol{H}}_a, \tilde{\boldsymbol{H}}_b) = \Phi_{\text{aff}}\big(\Phi(a,b)\big)$ *be the output of the Siamese cross-attention layer of the affinity module, let* $\boldsymbol{u}_a = [\text{meanpool}(\tilde{\boldsymbol{H}}_a);\ \text{maxpool}(\tilde{\boldsymbol{H}}_a)] \in \mathbb{R}^{2d}$*, where* $\text{meanpool}(\boldsymbol{X})_k = \frac{1}{n}\sum_i X_{ik}$ *and* $\text{maxpool}(\boldsymbol{X})_k = \max_i X_{ik}$ *pool over the* n *residue positions of a sequence (and analogously* $\boldsymbol{u}_b$*), and let the pair-level prediction be* $h(a,b) = \text{MLP}\big(c(\boldsymbol{u}_a, \boldsymbol{u}_b)\big)$ *with* $c(\boldsymbol{u},\boldsymbol{v}) = [\boldsymbol{u} \odot \boldsymbol{v};\ |\boldsymbol{u} - \boldsymbol{v}|]$*. Then* $h(a,b) = h(b,a)$ *for every pair of sequences. Moreover, if the residue-level prediction for residue* i *of* a *is* $g\big((\boldsymbol{H}_a^{(L)})_{i,:}\big)$ *for an MLP* g *applied row-wise, it is identical for the inputs* (a,b) *and* (b,a)*.*

The invariance is exact in exact arithmetic; in floating point it can be lost only through non-deterministic kernels or batch-dependent reduction order, and for all pairs we tested the outputs for (a,b) and (b,a) were bitwise identical (Appendix E).

The lemma, the proofs, and the reason the argument fails for concatenation-based models with global positions, are given in Appendix G.

Table 1: Performance metrics of various models evaluating accuracy, precision, recall, F_1-score, and AUPRC on the gold-standard PPI dataset (test set). Performance for the first seven architectures (Sledzieski et al., 2021; Ko et al., 2024) was reported by Reim et al. (2025), and performance for Topsy-Turvy (Singh et al., 2022) and PLM-interact were reported by Liu et al. (2025), and MINT by Ullanat et al. (2026). The total number of evaluation samples in the test set is 52,048. **NR**: Not reported.

Model	Accuracy	Precision	Recall	F1	AUPRC
RFC-40	0.577	0.629	0.374	0.496	0.584
2d-baseline	0.575	0.630	0.373	0.468	0.635
2d-Crossattention	0.641	0.660	0.589	0.623	0.678
2d-Selfattention	0.616	0.611	0.553	0.591	0.641
Richoux-ESM-2	0.633	0.627	0.654	0.640	0.655
D-SCRIPT-ESM-2	0.628	0.638	0.602	0.619	0.636
TUnA	0.645	**0.672**	0.580	0.622	0.692
PLM-interact	NR	0.63	**0.71**	0.66	0.69
Topsy-Turvy	NR	0.65	0.26	0.37	0.59
MINT	NR	NR	NR	NR	0.69
ESM-2 PPI	0.649	0.656	0.626	0.641	0.701
OMP	**0.665**	0.663	0.671	**0.667**	**0.712**

4 Experiments

We evaluate two models on three downstream tasks in which proteins exceeding 1,000 amino acids are frequent: OMP and, as the main baseline, the frozen ESM-2 (650M) backbone with the identical prediction head (denoted ESM-2 PPI). Both backbones stay frozen, and this is the closest ablation our compute allowed, although ESM-2 PPI also lacks the layer-wise cross-attention modules and has 325M fewer parameters. To limit data leakage, difficult to rule out entirely in biological data, we use three benchmarks with stringent cluster-based sequence separation: the gold-standard PPI dataset (52,048 test pairs), the TCR–epitope task of TDC-2 (five splits), and PINDER (residue-level interfaces, 220 test complexes), each with its published splits, so the prediction heads are trained and evaluated without leakage (Appendix B). Pre-training is a separate matter: it was not de-leaked against these benchmarks, as is also the case for the ESM-2 backbone of our baselines and for the STRING pre-training of MINT, and Appendix B.1 bounds the resulting pair-level overlap at 9.3% of the curated links.

4.1 Protein–protein interaction prediction

Binary PPI prediction evaluates whether a given pair of sequences physically binds. We use the third version of the gold-standard PPI dataset (Bernett, 2022), which minimizes leakage by sequence similarity and node degree (Appendix B.2). Truncation is consequential here: the chains above our 1,022-residue threshold amount to 76,119 occurrences in the training pairs, and cropping them removes an average of 736 amino acids each.

OMP achieves an accuracy of 66.5% and an AUPRC of 71.2% (Table 1), making it, to our knowledge, the first sequence-based model to reach this level on the benchmark. By contrast, ESM-2 PPI does not break the historical sequence-based plateau of 65% accuracy (Reim et al., 2025), and MINT is reported to reach an AUPRC of 0.69, close to PLM-interact and the PPI-specialized TUnA (Ullanat et al., 2026; Liu et al., 2025; Ko et al., 2024). OMP also achieves an AUROC of 0.7262 versus 0.7106 for ESM-2 PPI. A paired bootstrap test over 1,000 iterations on the full test set yields 99% confidence intervals for the gain of OMP over ESM-2 PPI of [0.0122, 0.0194] in AUROC and [0.0065, 0.0150] in AUPRC.

4.2 TCR–epitope binding prediction

We next evaluate whether frozen OMP embeddings transfer to a focused, domain-specific interaction task. We use the TCR–epitope benchmark of the Therapeutics Data Commons (TDC-2) (Velez-Arce et al., 2024) and select the split with randomly shuffled epitope and T-cell receptor (TCR) sequences (the "RN" set), the most challenging and biologically realistic scenario in the benchmark (Appendix B.3). The prediction head is the same as for PPI prediction, with a narrower MLP.

Table 2: AUROC of OMP and ESM-2 PPI (identical prediction head on frozen backbones) across the five splits of the TCR–epitope benchmark (RN set of TDC-2), with 99% paired bootstrap confidence intervals of the difference ($\Delta_{\text{OMP}-\text{ESM-2}}$). The strongest previously reported models are AVIB-TCR (0.576) (Velez-Arce et al., 2024) and MINT with its last layer fine-tuned (0.581) (Ullanat et al., 2026); the full TDC-2 baseline comparison is shown in Figure 7.

Split	N. Samples	OMP AUROC	ESM-2 PPI AUROC	99% CI (Diff)
0	71,915	0.6214	0.6131	[0.0025, 0.0139]
1	71,615	0.6186	0.6195	[-0.0067, 0.0045]
2	73,541	0.6209	0.5924	[0.0233, 0.0337]
3	70,236	0.5355	0.5334	[-0.0027, 0.0075]
4	71,023	0.6124	0.6109	[-0.0034, 0.0061]
Mean $\pm$ std	–	**0.6018** $\pm$ **0.0333**	0.5938 $\pm$ 0.0315	–

Table 3: Per-complex interface prediction on the cropped PINDER S test set ($n = 220$): median AUPRC and AUROC across complexes. OMP outperforms MINT (Wilcoxon $p = 0.0086$ for AUROC), whereas ESM-2 attains higher medians than OMP on both metrics (Wilcoxon $p = 0.0007$ for AUROC). Full distributions are shown in Figure 8.

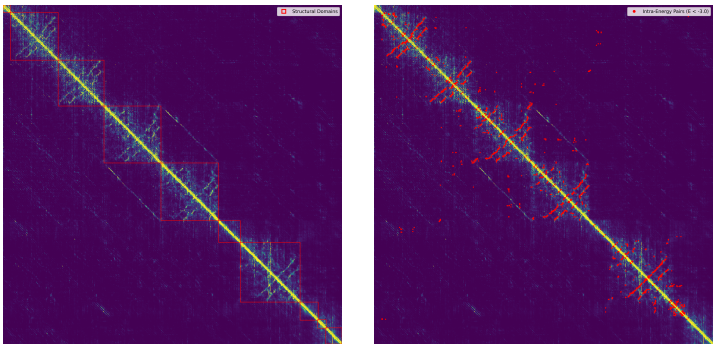
Model	Median AUPRC	Median AUROC
ESM-2	0.4049	0.7115
MINT	0.3442	0.6774
OMP	0.3770	0.6998

Among the models evaluated here, OMP is the only one with a mean AUROC above 60% (0.6018 $\pm$ 0.0333 across the five splits; Table 2 and Figure 7). ESM-2 PPI also performs strongly, with a mean AUROC of 0.5938, whereas MINT was reported to reach 0.581 with its last layer fine-tuned for this task (Ullanat et al., 2026). These are general-purpose and pair-level models, which isolates the effect of our pre-training but says nothing about the TCR-specific state of the art. On the closest published protocol, nine specialized predictors report AUROC between 0.501 and 0.573, none of them reaching 0.60, so the difficulty of this setting is not an artifact of general-purpose encoders; the figures and the protocol differences are in Appendix H. Because differences between the top models could be driven by stochastic variation, we computed 99% paired bootstrap confidence intervals over 1,000 iterations for each split: the gain of OMP in mean AUROC is small (0.008), is significant only in splits 0 and 2, and comes mostly from split 2. We report AUROC because it is the metric of the TDC-2 benchmark, which makes our models comparable to its published baselines, but the "RN" splits are not class-balanced, and under imbalance AUROC is the more forgiving metric; we do not report AUPRC here, which leaves precision on the minority class unmeasured.

4.3 Interface Prediction

Interface prediction classifies each residue of an interacting pair as participating or not in the interaction, using the native complex structure to define the ground truth (a residue is interface if one of its atoms lies within 10 Å of the partner chain). Because each complex has a variable number of residues, we compute AUPRC and AUROC per complex and compare their distributions across the test set. We use the PINDER S subset (Kovtun et al., 2024), whose clustering and de-leakage consider sequence identity and structural similarity, and which designates the longer chain of each pair as the receptor, another arbitrary ordering convention. Because the ground truth changes when a sequence is cropped and MINT is limited to 512 tokens, we randomly extracted a continuous window of 510 amino acids from each test complex; training and validation pairs used the maximum context window of each model. The window is not conditioned on the location of the interface, so pairs whose windows retain no interface residue in either chain are discarded, and the 220 complexes we evaluate are those that remain, each with at least one positive residue, so the per-complex AUPRC is always defined (Appendix B.4). Because the interface head is a per-residue MLP without cross-attention, we denote this baseline simply as ESM-2.

OMP significantly outperforms MINT, the concatenation-based interaction language model, in AUROC, whereas the single-sequence ESM-2 backbone attains higher median AUPRC and AUROC than OMP in this cropped setting (Table 3). We attribute this to the setting rather than to the backbone: cropping to the 512 residues per chain that MINT admits removes the long-range context OMP was pre-trained to exploit, and the residue-level head carries no cross-attention module; order



(a) Annotated structural domains. (b) Low intra-energy residue pairs.

Figure 2: Zero-shot raw self-attention of OMP at layer 30 for protein X, overlaid with **a** the annotated structural domains and **b** residue pairs of low intra-protein interaction energy ($E < -3.0\ kcal/mol$, via MM/GBSA (Wang et al., 2019)).

dependence cannot explain it, since both models are order-invariant here, unlike MINT. The attribution is not settled, however. The crop is imposed by MINT alone, and OMP and ESM-2 both admit 1,022 residues per chain, so the comparison could be run on full-length complexes; until it is, we cannot separate an intrinsic weakness on localized labels from an artifact of cutting the structural context.

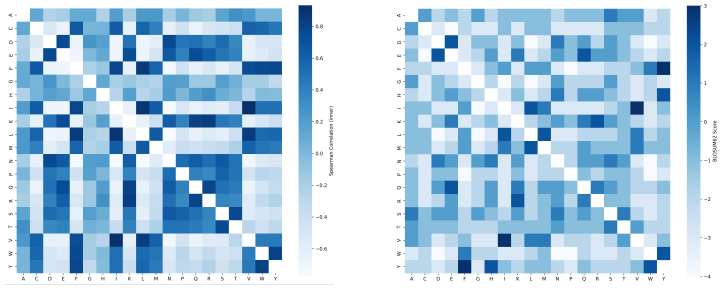
4.4 Interpretability analysis

We inspect attention maps extracted zero-shot, without fine-tuning, by taking for each layer the maximum attention weight over heads, and ask whether self-attention recovers structural domains, whether cross-attention concentrates on interacting regions, whether the backbone's biochemical knowledge survives warm-starting, and whether OMP develops attention sinks. We analyze an anonymized case study, protein X with its partner protein Y, whose identities are temporarily withheld to preserve author anonymity during double-blind review, and the public BCAR1–SRC complex (Table 8).

Structural domains. 86.01% of the strongest intra-protein residue interactions of protein X ($E < -3.0\ kcal/mol$, computed via MM/GBSA (Wang et al., 2019)) fall within the top 15% of the raw self-attention at layer 30 (Figure 2b), and 65.17% of the highest attention mass concentrates inside its annotated structural domains (Spearman correlation of 0.6295; Figure 2a), without any 3D structural supervision. Individual heads specialize: some attend within structural domains at medium range, others resolve short-range patterns along the diagonal, and a subset is unspecialized (Figure 6). The same co-localization holds for protein Y, analyzed in complex with protein X, and for the oncogenic complex between BCAR1 (p130Cas) and the kinase SRC (Barrett et al., 2013; Roskoski, 2015; Schuh et al., 2010; Nakamoto et al., 1996) (Figure 10). These three figures are measured on the anonymized pair, so they cannot be recomputed until its identity is disclosed; BCAR1–SRC is included precisely because it is public and reproduces the qualitative result, and the BLOSUM62 correspondence below and the attention sinks of Appendix F rest on public sequences alone.

Cross-attention. Qualitative cross-attention maps, including the interface of the BCAR1–SRC complex, are discussed in Appendix D.

Biochemical knowledge. Following Vig et al. (2021), every one of the 20 standard amino acids has a specialized self-attention head whose relative attention bias is at least twice its background frequency (Table 6). Residues that share a head (Ile–Val, Asn–Asp, Ser–Thr, Lys–Arg) have highly positive BLOSUM62 substitution scores (Henikoff & Henikoff, 1992), and head 28-14 specializes in aromatic residues. The Spearman correlation between the attention similarity matrix of OMP (Figure 3a) and BLOSUM62 (Figure 3b) is 0.748, nearly identical to the 0.749 of ESM-2 (650M); under the Pearson coefficient used by Vig et al. (2021), OMP reaches 0.79 and ESM-2 0.78, above every earlier PLM they report (Table 7). Warm-starting therefore preserved the biochemical knowledge of the self-attention parameters while pre-training integrated interaction constraints.



(a) Self-attention similarity between residues.

(b) Reference BLOSUM62 matrix.

Figure 3: BLOSUM substitution analysis of OMP self-attention. **a** Spearman correlation matrix between the layer- and head-wise self-attention profiles of each amino acid; negatively charged (Aspartic and Glutamic acid) and positively charged (Lysine and Arginine) residues exhibit strongly correlated attention profiles. Per-residue attention profiles are shown in Figure 9.

Attention sinks. Self-attention in OMP is not dominated by N-terminal boundary tokens, and hidden-state activations lack the massive outliers of large language models (Xiao et al., 2024; Sun et al., 2024); see Appendix F.

5 Conclusion

We presented OMP, a protein interaction language model that replaces sequence concatenation with Siamese cross-attention in parallel to self-attention in every layer of a pre-trained PLM. This design gives each chain the full context window of the backbone while preserving the same number of attention-score computations as concatenation, keeps intra- and inter-chain budgets independent, and makes the model order-invariant. With frozen embeddings, OMP improves over its ESM-2 backbone and prior sequence-based models on gold-standard PPI prediction and, more modestly, on TCR–epitope prediction, outperforms MINT on interface prediction, and exposes attention maps that recover protein domains without structural supervision, all with 4,032 GPU-hours over a fifth of the curated pairs: equipping a pre-trained PLM with pair-level structure does not require re-pretraining at the scale of the backbone. Pairing information during pre-training thus helps move past the plateau Reim et al. (2025) report for single-sequence encoders on leakage-free benchmarks.

Limitations and future work. First, the average gate openness remained below 1% (about 0.7% at the end of pre-training and still rising; Figure 4b), so cross-attention contributes a small fraction of the residual update on average; the gates are sparse rather than inactive, with openness concentrated in a few layers and saturating dimensions in 11 of the 33 (Appendix E). Whether that sparsity is architectural or a consequence of stopping early we cannot say: openness was still rising when the compute budget ran out, the same budget that prevented the ablations, which would require repeating the pre-training of a 975M-parameter model at 4,032 GPU-hours per run. Pre-training also covered only 19% of the curated pairs, because the sampler restarted at each weekly interval instead of resuming its state; whether the rest improves the results is untested, and it is the most direct extension of this work. Second, on interface prediction ESM-2 surpasses OMP in the cropped setting required for comparison with MINT. The crop is needed only for MINT, so the missing experiment is specific: OMP against ESM-2 on the full-length PINDER complexes, which both admit; until it is run we cannot separate an intrinsic weakness on localized labels from an artifact of the cropping. Third, the quantitative interpretability measurements come from the anonymized pair and cannot be verified until its identity is disclosed, although the qualitative behaviour reproduces on the public BCAR1–SRC complex. Fourth, pre-training was not de-leaked against the benchmarks: up to 9.3% of the curated links touch a family containing a gold-standard validation or test protein (Appendix B.1). The splits themselves are used as published, and the protein-level exposure is shared by the ESM-2 backbone of our baselines, which nevertheless remains at the plateau, so what stays specific to OMP is having seen interacting pairs. Finally, because OMP extends a PLM backbone rather than modifying it internally, the framework transfers to structure-aware (Hayes et al., 2025; Su et al., 2024; 2025; Liu et al., 2026), annotation-aware (Ko et al., 2026; Wu et al., 2024) and multimodal backbones such as LucaOne (He et al., 2025).

AI USE STATEMENT

We used a generative AI assistant (a large language model) for writing: text editing, improving the writing of the manuscript, and drafting the title and abstract from existing content. Generative AI was not used for synthetic data generation, dataset cleaning, hypothesis or methodology design, or model training. The authors reviewed all AI-assisted text and take full responsibility for the content of this work.

ETHICS STATEMENT

This work relies on publicly available data, including protein sequences and interactions from STRING, the gold-standard PPI dataset, TCR–epitope pairs from the Therapeutics Data Commons, and protein complex structures from PINDER. The identity of one protein pair used in the interpretability analysis is temporarily withheld to preserve author anonymity during double-blind review and will be disclosed after the review process. No new data were collected from human subjects.

REPRODUCIBILITY STATEMENT

The OMP architecture, including the cross-attention and gating formulation, is described in Section 3. Pre-training and fine-tuning hyperparameters, including the exact arguments of the best downstream configurations, are listed in Appendix A. Data sources and the curation, deduplication, cropping, and splitting procedures for STRING, the gold-standard PPI dataset, TDC-2, and PINDER are detailed in Appendix B. Anonymized source code and the pre-trained weights of the checkpoint used in all downstream tasks are available through the links in Appendix C, and the public protein sequences used in the interpretability case studies are listed in Table 8.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Chen Agassy, Bruria Samuel, Shahar Mayo, Asaf Zorea, and David Burstein. B-PPI: A Cross-Attention Model for Large-Scale Bacterial Protein-Protein Interaction Prediction. *bioRxiv*, 2025. doi: 10.64898/2025.12.23.696145.

Ramisa Alam, Sazan Mahbub, and Md Shamsuzzoha Bayzid. Pair-EGRET: enhancing the prediction of protein–protein interaction sites through graph attention networks and protein language models. *Bioinformatics*, 40(10):btae588, 2024. doi: 10.1093/bioinformatics/btae588.

Angela Barrett, Caroline Pellet-Many, Ian C Zachary, Ian M Evans, and Paul Frankel. p130Cas: a key signalling node in health and disease. *Cellular signalling*, 25(4):766–777, 2013. doi: 10.1016/j.cellsig.2012.12.019.

Judith Bernett. PPI prediction from sequence, gold standard dataset. Figshare dataset, version 3, 11 2022. URL https://figshare.com/articles/dataset/PPI_prediction_from_sequence_gold_standard_dataset/21591618/3.

Judith Bernett, David B Blumenthal, and Markus List. Cracking the black box of deep sequence-based protein–protein interaction prediction. *Briefings in Bioinformatics*, 25(2):bbae076, 03 2024. ISSN 1477-4054. doi: 10.1093/bib/bbae076.

Klára Briknarová, Fariborz Nasertorabi, Marnie L Havert, Ericka Eggleston, David W Hoyt, Chenglong Li, Arthur J Olson, Kristiina Vuori, and Kathryn R Ely. The serine-rich domain from Crk-associated substrate (p130cas) is a four-helix bundle. *Journal of Biological Chemistry*, 280 (23):21908–21914, 2005. doi: 10.1074/jbc.M501258200.

Thanh Hai Dang and Tien Anh Vu. xCAPT5: protein-protein interaction prediction using deep and wide multi-kernel pooling convolutional neural networks with protein language model. *BMC bioinformatics*, 25(1), 2024. doi: 10.1186/s12859-024-05725-6.

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, 2022. doi: 10.1109/TPAMI.2021.3095381.

Ahmed Essaghir, Nanda Kumar Sathiyamoorthy, Paul Smyth, Adrian Postelnicu, Stefan Ghiviriga, Alexandru Ghita, Anjana Singh, Shruti Kapil, Sanjay Phogat, and Gurpreet Singh. T-cell receptor specific protein language model for prediction and interpretation of epitope binding (ProtLM.TCR). *bioRxiv*, 2022. doi: 10.1101/2022.11.28.518167.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018.

Yong He, Pan Fang, Yongtao Shan, Yuanfei Pan, Yanhong Wei, Yichang Chen, Yihao Chen, Yi Liu, Zhenyu Zeng, Zhan Zhou, et al. Generalized biological foundation model with unified nucleic acid and protein language. *Nature Machine Intelligence*, pp. 1–12, 2025. doi: 10.1038/s42256-025-01044-4.

Steven Henikoff and Jorja G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992. doi: 10.1073/pnas.89.22.10915.

Young Su Ko, Jonathan Parkinson, Cong Liu, and Wei Wang. TUnA: an uncertainty-aware transformer model for sequence-based protein–protein interaction prediction. *Briefings in Bioinformatics*, 25(5):bbae359, 2024. doi: 10.1093/bib/bbae359.

Young Su Ko, Jonathan Parkinson, and Wei Wang. Scalable embedding fusion with protein language models: insights from benchmarking text-integrated representations. *Briefings in Bioinformatics*, 27(1):bbag014, 01 2026. ISSN 1477-4054. doi: 10.1093/bib/bbag014.

Daniel Kovtun, Mehmet Akdel, Alexander Goncearenco, Guoqing Zhou, Graham Holt, David Baugher, Dejun Lin, Yusuf Adeshina, Thomas Castiglione, Xiaoyun Wang, Céline Marquet, Matt McPartlon, Tomas Geffner, Gabriele Corso, Hannes Stärk, Zachary Carpenter, Emine Kucukbenli, Michael Bronstein, and Luca Naef. PINDER: The protein interaction dataset and evaluation resource. *bioRxiv*, 2024. doi: 10.1101/2024.07.17.603980.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Dan Liu, Francesca Young, Kieran D Lamb, Adalberto Claudio Quiros, Alexandrina Pancheva, Crispin J Miller, Craig Macdonald, David L Robertson, and Ke Yuan. PLM-interact: extending protein language models to predict protein-protein interactions. *Nature Communications*, 16(1): 9012, 2025. doi: 10.1038/s41467-025-64512-w.

Jun Liu, Hungyu Chen, and Yang Zhang. A paired sequence language model for protein-protein interaction modeling. *Nature Communications*, 2026. doi: 10.1038/s41467-026-70457-5.

Peter D Mace, Yann Wallez, Małgorzata K Dobaczewska, JeongEun J Lee, Howard Robinson, Elena B Pasquale, and Stefan J Riedl. NSP-Cas protein structures reveal a promiscuous interaction module in cell signaling. *Nature structural & molecular biology*, 18(12):1381–1387, 2011. doi: 10.1038/nsmb.2152.

Cyril Malbranke, Gionata Paolo Zalaffi, and Anne-Florence Bitbol. ProteomeLM: A proteome-scale language model enables accurate and rapid prediction of protein–protein interactions and gene essentiality across taxa. *Proceedings of the National Academy of Sciences*, 123(21):e2524201123, 2026. doi: 10.1073/pnas.2524201123.

Tetsuya Nakamoto, Ryuichi Sakai, Keiya Ozawa, Yoshio Yazaki, and Hisamaru Hirai. Direct Binding of C-terminal Region of p130 to SH2 and SH3 Domains of Src Kinase. *Journal of Biological Chemistry*, 271(15):8959–8965, 1996. ISSN 0021-9258. doi: 10.1074/jbc.271.15.8959.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *BioRxiv*, 2025. doi: 10.1101/2025.06.14.659707.

Cong Qi, Hanzhang Fang, Siqi Jiang, Tianxing Hu, and Zhi Wei. LANTERN: TCR-peptide binding prediction via large language model representations. *PeerJ*, 14:e20980, 2026. doi: 10.7717/peerj.20980.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. doi: 10.52202/085713-3345. URL https://openreview.net/forum?id=1b7whO4SfY.

Timo Reim, Anne Hartebrodt, David B Blumenthal, Judith Bernett, and Markus List. Deep learning models for unbiased sequence-based PPI prediction plateau at an accuracy of 0.65. *Bioinformatics*, 41(Supplement_1):i590–i598, 07 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf192.

Robert Roskoski. Src protein-tyrosine kinase structure, mechanism, and small molecule inhibitors. *Pharmacological Research*, 94:9–25, 2015. ISSN 1043-6618. doi: 10.1016/j.phrs.2015.01.003.

Natasha R Schuh, Michael S Guerrero, Randy S Schrecengost, and Amy H Bouton. BCAR3 regulates Src/p130Cas association, Src kinase activity, and breast cancer adhesion signaling. *Journal of Biological Chemistry*, 285(4):2309–2317, 2010. doi: 10.1074/jbc.M109.046631.

Rohit Singh, Kapil Devkota, Samuel Sledzieski, Bonnie Berger, and Lenore Cowen. Topsy-Turvy: integrating a global view into sequence-based PPI prediction. *Bioinformatics*, 38(Supplement_1): i264–i272, 06 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac258.

Samuel Sledzieski, Rohit Singh, Lenore Cowen, and Bonnie Berger. D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Systems*, 12(10):969–982, 2021. doi: 10.1016/j.cels.2021.08.010.

Samuel Sledzieski, Meghana Kshirsagar, Minkyung Baek, Rahul Dodhia, Juan Lavista Ferres, and Bonnie Berger. Democratizing protein language models with parameter-efficient fine-tuning. *Proceedings of the National Academy of Sciences*, 121(26):e2405840121, 2024. doi: 10.1073/pnas.2405840121.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein Language Modeling with Structure-aware Vocabulary. In *ICLR*, 2024. URL https://openreview.net/forum?id=6MRm3G4NiU.

Jin Su, Zhikai Li, Tianli Tao, Chenchen Han, Yan He, Fengyuan Dai, Qingyan Yuan, Yuan Gao, Tong Si, Xuting Zhang, et al. Democratizing protein language model training, sharing and collaboration. *Nature Biotechnology*, pp. 1–7, 2025. doi: 10.1038/s41587-025-02859-7.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023. doi: 10.1093/nar/gkac1000.

Varun Ullanat, Bowen Jing, Samuel Sledzieski, and Bonnie Berger. Learning the language of protein-protein interactions. *Nature Communications*, 2026. doi: 10.1038/s41467-025-67971-3.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. doi: 10.48550/arXiv.1706.03762.

Alejandro Velez-Arce, Kexin Huang, Michelle M Li, Xiang Lin, Wenhao Gao, Bradley Pentelute, Tianfan Fu, Manolis Kellis, and Marinka Zitnik. Signals in the Cells: Multimodal and Contextualized Machine Learning Foundations for Therapeutics. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*, 2024. URL https://openreview.net/forum?id=kL8dlYp6IM.

Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Rajani. BERTology Meets Biology: Interpreting Attention in Protein Language Models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YWtLZvLmud7.

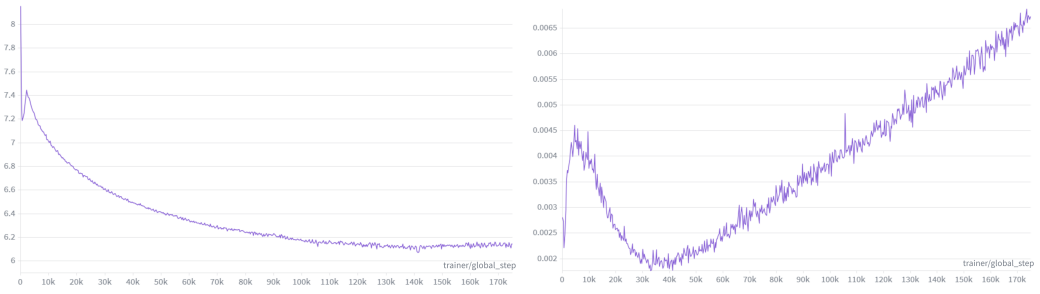
Ercheng Wang, Huiyong Sun, Junmei Wang, Zhe Wang, Hui Liu, John Z. H. Zhang, and Tingjun Hou. End-point binding free energy calculation with mm/pbsa and mm/gbsa: Strategies and applications in drug design. *Chemical Reviews*, 119(16):9478–9508, 2019. doi: 10.1021/acs.chemrev.9b00055. PMID: 31244000.

Kevin E. Wu, Howard Chang, and James Zou. ProteinCLIP: enhancing protein language models with natural language. *bioRxiv*, 2024. doi: 10.1101/2024.05.14.594226.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

A Pre-training and fine-tuning details

A.1 Pre-training of OMP



(a) Perplexity during pre-training.

(b) Gate openness during pre-training.

Figure 4: Pre-training metrics of OMP. **a–b** Line graphs tracking perplexity and gate openness during OMP pre-training; perplexity was evaluated over 32,000 validation pairs, and gate openness was computed as the average across gate dimensions per token per sequence for the first 20 validation batches.

The backbone is pre-LayerNorm, and both branches read the same layer input and are summed into the residual stream in a single step. Writing LN_s, LN_c and LN_f for the self-attention, cross-attention and feed-forward normalizations, each applied before its module, the update of chain a with partner b is

$$\begin{aligned} \boldsymbol{U}_a &= \boldsymbol{H}_a^{(\ell-1)} + \mathrm{SelfAttn}\big(\mathrm{LN}_s(\boldsymbol{H}_a^{(\ell-1)})\big) + \mathrm{GLU}\Big(\mathrm{CrossAttn}\big(\mathrm{LN}_c(\boldsymbol{H}_a^{(\ell-1)}),\, \mathrm{LN}_c(\boldsymbol{H}_b^{(\ell-1)})\big)\Big), \\ \boldsymbol{H}_a^{(\ell)} &= \boldsymbol{U}_a + \mathrm{FFN}\big(\mathrm{LN}_f(\boldsymbol{U}_a)\big), \end{aligned} \tag{5}$$

and symmetrically for b. The cross-attention branch does not reuse the normalization of the self-attention branch: it applies its own layer normalization, with its own parameters, to the same layer input, takes its queries from chain a and its keys and values from chain b, and its output passes through the GLU before entering the residual stream. Every module and normalization is shared between the two chains.

Each of the 33 layers holds, in the released checkpoint, the following modules: `self_attn` with its four 1280×1280 projections and the rotary embedding, `self_attn_layer_norm`, `cross_attn` with its own four projections and no rotary embedding, `cross_attn_layer_norm`, `gate_dense` of shape 2560×1280 for the GLU of Eq. (4), `fc1` and `fc2` of the feed-forward network, and `final_layer_norm`. The parameter names of the published weights can be checked against this list.

The masked-token prediction head is the one used to pre-train ESM-2 (the `RobertaLMHead` class), which consists of a dense layer, a Gaussian Error Linear Unit (GELU) activation, layer normalization (the `ESM1bLayerNorm` class), and a projection to the vocabulary.

Our pre-training setup is similar to that of MINT and ESM-2 (Ullanat et al., 2026; Lin et al., 2023). Transfer learning was initiated by warm-starting the self-attention and feed-forward network modules with pre-trained ESM-2 weights. To prevent catastrophic forgetting, the gates of the newly introduced GLU modules were initialized nearly closed (bias -6), and the learning rate underwent a warm-up phase for the first 2,000 global steps. Following warm-up, the learning rate peaked at 0.0004 before linearly decaying to 10% of its maximum value by the conclusion of pre-training. We used the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-8}$), with an L2 weight decay of 0.01. Incorporating the 650M-parameter ESM-2 model as a core backbone, OMP encompasses a total of 975 million trainable parameters during pre-training. Writing d for the hidden size, $4d$ for the feed-forward size and $n = n_a + n_b$ for the total number of residues, a concatenation-based layer spends $4nd^2$ on the attention projections, $2n^2d$ on the score and context products, and $8nd^2$ on the feed-forward network. OMP spends $8nd^2$ on projections, because self- and cross-attention each project their own queries, keys and values, a further $2nd^2$ on the gate of the GLU, the same $2n^2d$ on scores and contexts since $n_a^2 + n_b^2 + 2n_an_b = n^2$, and the same $8nd^2$ on the feed-forward network. The matrix-multiplication cost of a layer therefore grows from $12nd^2$ to $18nd^2$, a factor of 1.5 at equal total length that matches the ratio of parameters, and the same factor applies to the backward pass. The quadratic term is unchanged, so the relative overhead decreases as the pair gets longer, which is the regime the model is built for.

The published weights contain 975,827,009 parameters, of which the cross-attention machinery accounts for 324,741,120: per layer, four 1280×1280 projections with biases (6,558,720), the 2560×1280 gate of the GLU with its bias (3,279,360) and the cross-attention layer normalization (2,560), that is 9,840,640 across each of the 33 layers. The remaining 651,085,889 parameters are those of the warm-started ESM-2 backbone, so the entire increase is attributable to the parallel modules and none of their weights is shared with self-attention.

Prior to the forward pass, we randomly sample 15% of the tokens from each input sequence. These selected tokens are modified as follows: 80% are replaced with the `<mask>` token, 10% are replaced with tokens sampled from a uniform distribution, and the remaining 10% are left unchanged. For evaluation, we compute the validation perplexity, defined as the exponential of the cross-entropy loss between the predictions of OMP and the true tokens.

The model was trained on 6 NVIDIA H100 GPUs in bf16 mixed precision, with a per-GPU batch size of 2 and 43 gradient accumulation steps, for an effective batch size of 516. Training ran in one-week intervals, and the checkpoint used in all downstream tasks was obtained at the end of the fourth interval, for a total of 4,032 GPU-hours.

OMP reached a validation perplexity of 6.0 on held-out STRING pairs (Figure 4a), converging after four epochs over 18,421,200 pairs. For reference, ESM-2 15B reports a validation perplexity of 6.4 after training on approximately 65 million unique sequences over at least 15 epochs (Lin et al., 2023); because the validation sets differ, these values are not directly comparable.

A.2 FINE-TUNING OF OMP

Identical to the pre-training configuration, we used the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-8}$), with an L2 weight decay of 0.01. The freeze_percentage parameter was set to 1.0 on all runs to use exclusively frozen embeddings. With LN'_c and LN'_f its two normalizations,

$$\boldsymbol{V}_a = \boldsymbol{H}_a^{(L)} + \mathrm{CrossAttn}'\big(\mathrm{LN}'_c(\boldsymbol{H}_a^{(L)}),\, \mathrm{LN}'_c(\boldsymbol{H}_b^{(L)})\big), \qquad \tilde{\boldsymbol{H}}_a = \boldsymbol{V}_a + \mathrm{FFN}'\big(\mathrm{LN}'_f(\boldsymbol{V}_a)\big), \quad (6)$$

and symmetrically for b, with all parameters shared by the two chains and no positional encoding, as in the cross-attention of the backbone.

Training a head is comparatively cheap: the longest run took 21 hours on the gold-standard PPI dataset, 4 hours on PINDER and 1.5 hours on TCR–epitope, and the best epoch was usually the first. The results reported were obtained using the following configurations.

For the gold-standard PPI prediction the affinity module was trained (refer to Figure 5). The arguments used to run the best configuration are: `--batch_size 32 --accumulate_grad 2 --epochs 20 --mlp_dropout 0.2 --mlp_hidden 5120 --lr 0.0002 --head_type deep --add_light_cross`.

For the domain-specific TCR epitope interaction prediction we also trained the affinity module (refer to Figure 5). The arguments used to run the best configuration are: `--batch_size 64 --accumulate_grad 1 --epochs 5 --mlp_dropout 0.2 --mlp_hidden 2560 --lr 0.0002 --head_type deep --add_light_cross`.

In both pair-level tasks, the affinity module includes the Siamese cross-attention layer of Figure 5 (`--add_light_cross`); the interface head does not. PLM per-token embeddings should carry enriched information for token-level tasks such as interface prediction. Therefore, we attached an MLP head (without cross-attention) to each encoder in order to compare raw frozen embeddings. The arguments used to run the best configuration are: `--batch_size 32 --accumulate_grad 2 --epochs 5 --mlp_dropout 0.2 --mlp_hidden 1280 --lr 0.0001`.

B DATA PREPARATION

B.1 USAGE OF THE STRING DATASET

We compiled over 59.3 million unique sequences and 2.3 billion interactions from the STRING database (Szklarczyk et al., 2023). To enhance data diversity per batch and prevent leakage between the pre-training and validation splits, we followed a curation protocol similar to MINT (Ullanat et al., 2026) and clustered the sequences into 15.5 million families using MMseqs2 with a 50% sequence similarity threshold. This rigorous separation also mitigates the risk of overfitting given our model's large parameter footprint. For each pair of families connected by at least one link, a single representative interaction was selected uniformly at random. This filtering pipeline yielded 382,108,560 interactions across 29,024,964 unique sequences, from which 250,000 pairs were randomly reserved for validation. The final training set was isolated by removing any sequence families present in the validation set, resulting in 96,420,565 interactions across 16,392,879 unique sequences. Because validation pairs are sampled uniformly over links, they are concentrated in highly connected families, so removing every training link that touches one of these families discards a disproportionate share of interactions. Since some sequences exceed our context window of 1022 tokens, we perform on-the-fly cropping allowing a single pair to be learned under different contexts.

Overlap with the downstream benchmarks. The clustering above separates the pre-training and validation splits of STRING, but it does not remove proteins or pairs that occur in the test sets of the downstream benchmarks, and the checkpoint reported here was pre-trained without such a filter; the positives of the gold-standard benchmark derive from HIPPIE, so some of them are also STRING links. To quantify the resulting exposure, we added the validation and test proteins of the gold-standard dataset to the sequence collection and repeated the MMseqs2 clustering at 50% identity. The number of families decreased from 15,556,389 to 15,554,053: the benchmark proteins formed no new families and merged existing ones, so essentially every test protein has a relative above 50% identity among the STRING sequences. Overlap of this kind is unavoidable for any PLM pre-trained at proteome scale and is shared by ESM-2, and therefore by our ESM-2 baseline and by

MINT, which is in addition pre-trained on the same STRING interactions (Ullanat et al., 2026); the comparison between the three models is therefore not distorted by it. Two further points bound the concern. The benchmarks are used with their published splits, so the prediction heads are trained and evaluated under the leakage controls their authors designed; and ESM-2 PPI, which shares the protein-level exposure, does not surpass the accuracy plateau, so familiarity with the test proteins does not by itself account for the result. What is specific to a pair-level model is the overlap of interactions, which is what we quantify next. Recomputing the representative links over the merged families yields 381,446,027 interactions, of which 35,298,625 (9.3%) touch a family that contains a validation or test protein; reserving them leaves 346,147,402 interactions. We report 9.3% as an upper bound on the pair-level exposure of the reported checkpoint: the filter discards every link of an affected family, and pre-training covered a subset of 18.4M pairs whose intersection with the reserved set we did not measure. Pre-training on the filtered set is left for future work.

We downloaded the files `protein.sequences.v12.0.fa.gz` (sequences) and `protein.physical.links.full.v12.0.txt.gz` (interactions) from the STRING download page (Szklarczyk et al., 2023). Pre-trained ESM-2 weights were obtained through `torch.hub.load` in the PyTorch library.

B.2 Usage of the gold-standard PPI dataset

The gold-standard PPI dataset (Bernett, 2022; Bernett et al., 2024) was designed to minimize data leakage by sequence similarity and node degree. The authors first partitioned the proteome into three sets while minimizing sequence similarity between them, by minimizing the cut size in a sequence similarity graph, which resulted in the sets P_0 (6,987 proteins), P_1 (6,987), and P_2 (6,379). They then mapped interactions from the HIPPIE database (version 2.3) into each set, keeping only intra-set pairs, and generated negative pairs by sampling proteins weighted by their degree until the sets were balanced. After removing redundant samples, the published sets are Intra_0 with 59,260 PPIs (validation), Intra_1 with 163,192 PPIs (training), and Intra_2 with 52,048 PPIs (test).

To ensure a fair evaluation between baselines, all training, validation, and testing splits were tokenized and cropped prior to fine-tuning. All pairs were correctly processed.

The gold-standard PPI dataset (Bernett, 2022; Bernett et al., 2024) Version 3 was downloaded from PPI prediction from sequence, gold standard dataset in figshare.

B.3 Usage of the Therapeutics Data Commons dataset

The Therapeutics Data Commons dataset (TDC-2) presents datasets, tools, models, and benchmarks focused on AI for drug discovery (Velez-Arce et al., 2024). In particular, we used the "TDC.tchard" dataset from TDC-2 (the "TCREpitopeGroup" benchmark in the code of the Python library published by TDC-2), which was created for T-cell receptor (TCR) and epitope interaction prediction, with an emphasis on assessing the capability of a model to effectively generalize to unseen TCRs and epitopes.

We selected the randomly shuffled epitope and T-cell receptor (TCR) sequence setup (the "RN" set), which represents the most challenging and biologically realistic evaluation scenario within the benchmark. To comprehensively assess generalization, the dataset is divided into five train–test splits: split 0 (522,076 train / 71,915 test), split 1 (522,253 / 71,615), split 2 (520,403 / 73,541), split 3 (523,618 / 70,236), and split 4 (522,847 / 71,023).

To ensure a fair evaluation between baselines, all training and test splits were tokenized and cropped prior to fine-tuning. The sizes that were reported are after the correct deduplication, cropping and tokenization of the sequences.

The TCR–epitope "RN" set from Therapeutics Data Commons dataset (TDC-2) (Velez-Arce et al., 2024) was prepared using the "TCREpitopeGroup" Python class from the PyTDC Python library.

B.4 Usage of the PINDER dataset

The PINDER dataset (Kovtun et al., 2024) was carefully designed to compile structural data for PPI complex modeling and docking. Within PINDER, clustering and de-leakage procedures are exe-

cuted by evaluating both primary sequence identity and 3D structural similarity, making it suitable for training and evaluating deep learning models.

PINDER is too small to pre-train an interaction language model, but it suffices to train a lightweight residue-level head, and its structural curation enables a rigorous, leakage-free evaluation of interface prediction. Interface labels are obtained with PINDER's `get_interface_residues` on the native dimer, which marks residues with an atomic contact within 10 Å of the partner chain. We evaluate on the PINDER S subset, which contains only heterodimers, is diverse at the family level, and excludes complexes predicted by AlphaFold2. We use its *holo* structures: binary PPIs that were identified between pairs of proteins with at least one pair of backbone atoms within a 10 Å threshold.

Cropping is applied once, during data preparation, and the tokenized windows are cached and reused by every model, so all systems are evaluated on identical inputs. For a chain longer than the context limit, the start of the window is drawn uniformly at random and is not conditioned on where the interface lies; the labels are then read off the native dimer within that window. Because a window can therefore exclude the interface entirely, a pair is discarded when neither of its two cropped chains retains an interface residue, and the metrics are computed per complex over both chains together, so every evaluated complex contributes at least one positive. The discarding is a property of the cropped data, not of any model, and applies equally to OMP, MINT and ESM-2.

To ensure a fair evaluation between baselines, all training and test splits were tokenized and cropped prior to fine-tuning, resulting in 41,739 pairs kept for training and 1,955 for testing. We also used PINDER S for testing (a subset of 220 pairs post-processing), as it contains strictly heterodimers, exhibits high diversity at the family level, and excludes complexes predicted by AlphaFold2.

PINDER dataset (Kovtun et al., 2024) splits were prepared using the PINDER Python library.

C Code and parameter availability

Our anonymized source code can be found at `https://anonymous.4open.science/r/OMP`, and the de-anonymized repository will be released upon acceptance.

The pre-trained weights for the checkpoint used in downstream tasks can be found at `https://huggingface.co/OMP-ICLR/OMP` or downloaded via the Hugging Face Hub Python library using "OMP-ICLR/OMP" as repository id. For usage, safetensors library must be used, and the code might need adaptation from using ".ckpt" to ".pt" files. The weights will be released under the authors' names upon acceptance.

Additional software used in this project: MMseqs2 for sequence clustering. MM/GBSA (molecular mechanics generalized Born surface area) for intra-protein energy calculation.

D Cross-attention interpretability

Unlike the block-diagonal patterns of self-attention, cross-attention maps are dominated by vertical and horizontal lines, indicating residues that act as sequence-wide hubs broadcasting information to the partner chain; in intermediate layers, cross-attention mirrors the patterns of self-attention (Figure 11). Although contact zones are harder to resolve through unsupervised training (Reim et al., 2025), cross-attention at layer 20 targets the known interface of the BCAR1–SRC complex (Figure 12b): residues of BCAR1 query the SH3 domain of SRC (residues 84–145) (Roskoski, 2015), consistent with the requirement of the SH3-binding motifs of BCAR1 for kinase association (Schuh et al., 2010; Nakamoto et al., 1996). Multiplying the attention matrices by the corresponding gate openness to obtain a gated cross-attention (Figure 13b) yields a sharper map at the adjacent layer 21: the SH3 domain of SRC is routed to the SH3-binding motif "RPLPSPPKF" of BCAR1 (residues 635–643) (Nakamoto et al., 1996; Barrett et al., 2013; Briknarová et al., 2005).

E Empirical order invariance and gate openness

We complement Section 3.5 with measurements on released checkpoints. We sampled 60 reviewed human sequences of 300–1,000 residues from UniProt, formed 30 disjoint pairs, and ran every pair in both orders.

Order invariance. For OMP we compared the final-layer representations of each chain. For all 30 pairs the two orders give bitwise identical representations, $\max_{i,k} |(\boldsymbol{H}_a^{(L)})_{ik}(a,b) - (\boldsymbol{H}_a^{(L)})_{ik}(b,a)| = 0$, with entries of magnitude up to 7.6, as predicted by Lemma 1. For MINT we used the released pre-trained checkpoint together with its released prediction head trained on the gold-standard PPI dataset (Ullanat et al., 2026), cropping each chain to 510 residues. Its encoder representations are also preserved up to numerical error ($\max|\Delta| \leq 8 \times 10^{-6}$ per token, for entries of magnitude up to 4.7), but its head concatenates the two chain embeddings in the input order and is therefore asymmetric: swapping the inputs changes the predicted interaction probability by 0.248 on average and flips the predicted label for 16 of the 30 pairs (Table 4). In concatenation-based pipelines, order dependence therefore need not originate in the encoder; here it originates in the pair representation passed to the classifier. OMP is invariant end to end because both its encoder (Lemma 1) and its prediction head (Theorem 1) are symmetric.

Table 4: Effect of swapping the two input proteins on 30 human protein pairs. Encoder drift is the maximum absolute change of a chain's final-layer representation; Δp is the change in predicted interaction probability, and flips count pairs whose predicted label changes at the 0.5 threshold. OMP predictions are invariant by Theorem 1 because its head is symmetric.

Model	Encoder drift	Mean $\lvert\Delta p\rvert$	Median $\lvert\Delta p\rvert$	Max $\lvert\Delta p\rvert$	Label flips
OMP	0 (exact)	invariant by construction (Theorem 1)			
MINT	8×10^{-6}	0.248	0.194	0.565	16/30

Gate openness. The same forward passes record the openness $\sigma(\boldsymbol{g})$ of every GLU of the released OMP checkpoint. Averaged over tokens, dimensions, and layers it is 0.005, but it is far from uniform: 19 of the 33 layers have mean openness below 0.001, 8 layers exceed 0.01, and 11 layers contain dimensions that saturate ($\sigma(\boldsymbol{g}) > 0.9$). The last layer is the most active, followed by layers 2, 9, and 29 (Table 5). The cross-attention pathway is therefore sparse rather than inactive: a small subset of layers and dimensions carries it, which is consistent with the sparsity expected of gated attention (Qiu et al., 2025) and with the low average openness reported in Figure 4b.

Table 5: Gate openness $\sigma(\boldsymbol{g})$ of the released OMP checkpoint, measured over 30 protein pairs: the four most active layers and the aggregate over all 33 layers.

Layer	Mean	99th pct.	Max	% dims > 0.1
33	0.023	0.279	1.000	4.84
2	0.019	0.074	0.995	0.42
9	0.014	0.053	0.719	0.10
29	0.013	0.080	0.998	0.62
All layers	0.005	–	1.000	–
Median layer	0.0001	–	–	–

F ATTENTION SINKS AND MASSIVE ACTIVATIONS

This appendix asks whether OMP shows the attention pathologies reported for large transformers. An attention sink is a bias toward the first tokens of a sequence, observed in many transformer models (Xiao et al., 2024): because the softmax distributes a fixed amount of attention over the sequence, a model can place surplus attention on the initial positions regardless of their relevance, usually on delimiter tokens such as `<cls>` or `<eos>`.

This matters for proteins because many sequences begin with a signal peptide, a short N-terminal segment that determines sub-cellular localization. A PPI model can legitimately use it to rule out an interaction between proteins of different compartments, but a head that is dominated by those first residues is reading localization rather than binding.

Figure 14 shows the attention received by each position in a forward pass over the SRC–BCAR1 complex. Self- and cross-attention have similar distribution shapes, and self-attention shows no N-terminal bias. Gating has been reported to reduce attention sinks (Qiu et al., 2025), which is one reason our architecture includes GLUs, so we measured the gated cross-attention as well.

The GLU flattens the background of the distribution but does not remove the bias on the boundary tokens themselves. We read this residual bias as a sign that the cross-attention modules are less

trained than the self-attention ones, which were warm-started from ESM-2, although we have not tested that explanation. The peaks of self-attention also appear in cross-attention, which is consistent with some cross-attention layers reproducing the positional pattern of self-attention (Figure 11).

Attention sinks do not require large attention weights (Qiu et al., 2025), so we also measured the magnitude of the hidden states in the residual stream. Outlier activations appear in both chains (Figure 15), but the distribution has no heavy tail: the ratio of the maximum to the median activation is about $190\times$, against the $15{,}000\times$ reported for LLaMA2-70B (Sun et al., 2024).

G Proof of order invariance

We now formalize why OMP is order-invariant by construction: its predictions for a pair do not depend on which protein is given first. Let a and b be amino acid sequences, and let $\boldsymbol{H}_a^{(\ell)} \in \mathbb{R}^{n_a \times d}$ and $\boldsymbol{H}_b^{(\ell)} \in \mathbb{R}^{n_b \times d}$ be their hidden states after layer ℓ, whose rows include the `<cls>` and `<eos>` tokens that each chain receives separately. We consider the model at inference, with dropout disabled, and rely on the following structural properties of OMP: (i) the embedding layer E is applied to each sequence separately, $\boldsymbol{H}_a^{(0)} = E(a)$ and $\boldsymbol{H}_b^{(0)} = E(b)$, with no chain-identity embedding; (ii) RoPE is applied only in self-attention, with positions restarting in each chain, and cross-attention uses no positional encoding (Section 3.2); (iii) all parameters are shared by both chains; (iv) no attention mask distinguishes the two chains, padding positions are masked in self- and cross-attention, and pooling is restricted to the residue positions of each sequence, excluding padding and the special tokens, so that the hidden states and pooled vectors of a sequence do not depend on the padding of its batch slot; and (v) both chains are updated simultaneously, with the keys and values of each cross-attention computed from the partner's layer-$(\ell-1)$ state. Under these properties, every layer $\ell = 1, \dots, L$ applies a single function F_ℓ to each chain and its partner:

$$\boldsymbol{H}_a^{(\ell)} = F_\ell\big(\boldsymbol{H}_a^{(\ell-1)}, \boldsymbol{H}_b^{(\ell-1)}\big), \qquad \boldsymbol{H}_b^{(\ell)} = F_\ell\big(\boldsymbol{H}_b^{(\ell-1)}, \boldsymbol{H}_a^{(\ell-1)}\big), \tag{7}$$

where F_ℓ is the update of Eq. (5), comprising self-attention on $\boldsymbol{X}$, cross-attention from $\boldsymbol{X}$ to $\boldsymbol{Y}$ (Eq. (3)), the gated linear unit, the residual connections, the three layer normalizations, and the feed-forward network. We write $\Phi(a, b) = (\boldsymbol{H}_a^{(L)}, \boldsymbol{H}_b^{(L)})$ for the output of the encoder and $\tau(\boldsymbol{X}, \boldsymbol{Y}) = (\boldsymbol{Y}, \boldsymbol{X})$ for the swap of a pair.

Lemma 1 (Swap equivariance of the encoder)**.** *Under properties (i)–(v), for every pair of sequences a, b, $\Phi(b, a) = \tau\big(\Phi(a, b)\big)$.*

Proof. Let $(\boldsymbol{H}_a^{(\ell)}, \boldsymbol{H}_b^{(\ell)})$ be the hidden states for the input (a, b) and $(\boldsymbol{G}_1^{(\ell)}, \boldsymbol{G}_2^{(\ell)})$ those for the input (b, a). We show by induction on ℓ that $\boldsymbol{G}_1^{(\ell)} = \boldsymbol{H}_b^{(\ell)}$ and $\boldsymbol{G}_2^{(\ell)} = \boldsymbol{H}_a^{(\ell)}$. For $\ell = 0$, $\boldsymbol{G}_1^{(0)} = E(b) = \boldsymbol{H}_b^{(0)}$ and $\boldsymbol{G}_2^{(0)} = E(a) = \boldsymbol{H}_a^{(0)}$. If the claim holds for $\ell - 1$, then by Eq. (7), $\boldsymbol{G}_1^{(\ell)} = F_\ell(\boldsymbol{G}_1^{(\ell-1)}, \boldsymbol{G}_2^{(\ell-1)}) = F_\ell(\boldsymbol{H}_b^{(\ell-1)}, \boldsymbol{H}_a^{(\ell-1)}) = \boldsymbol{H}_b^{(\ell)}$ and, likewise, $\boldsymbol{G}_2^{(\ell)} = F_\ell(\boldsymbol{H}_a^{(\ell-1)}, \boldsymbol{H}_b^{(\ell-1)}) = \boldsymbol{H}_a^{(\ell)}$. Taking $\ell = L$ proves the lemma. □

We restate Theorem 1 in the notation above and prove it.

Proof of Theorem 1. By Lemma 1, $\Phi(b, a) = \tau\big(\Phi(a, b)\big)$. The Siamese block of the affinity module (Figure 5), used for both PPI and TCR–epitope prediction, is not a copy of the backbone layer, but Eq. (6) applies one shared map to $(\boldsymbol{X}, \boldsymbol{Y})$ and to $(\boldsymbol{Y}, \boldsymbol{X})$, with shared parameters and no positional encoding, which is all the inductive step of Lemma 1 requires; it therefore gives $\Phi_{\text{aff}}\big(\tau(\boldsymbol{X}, \boldsymbol{Y})\big) = \tau\big(\Phi_{\text{aff}}(\boldsymbol{X}, \boldsymbol{Y})\big)$; hence $\Phi_{\text{aff}}\big(\Phi(b, a)\big) = (\tilde{\boldsymbol{H}}_b, \tilde{\boldsymbol{H}}_a)$. Pooling acts on each chain separately, so the input (b, a) yields the pooled pair $(\boldsymbol{u}_b, \boldsymbol{u}_a)$. Because the element-wise product is commutative and $|\boldsymbol{u}_b - \boldsymbol{u}_a| = |\boldsymbol{u}_a - \boldsymbol{u}_b|$, we have $c(\boldsymbol{u}_b, \boldsymbol{u}_a) = c(\boldsymbol{u}_a, \boldsymbol{u}_b)$; the MLP therefore receives identical inputs and $h(b, a) = h(a, b)$. For residue-level predictions, g acts on each row of $\boldsymbol{H}_a^{(L)}$, which by Lemma 1 is identical for the inputs (a, b) and (b, a). □

The invariance is exact in exact arithmetic; in floating point it can be lost only through non-deterministic kernels or batch-dependent reduction order, and for all pairs we tested the outputs for (a, b) and (b, a) were bitwise identical (Appendix E).

Remark 1. The argument does not hold for concatenation-based models with global positions, such as MINT. With input $[a; b]$ and rotary positional embeddings over the concatenated sequence, the attention logit between residue i of a and residue j of b depends on the relative position $i-(n_a+j)$, whereas for input $[b; a]$ it depends on $(n_b + i) - j$. These offsets differ by $n_a + n_b$, so the logits, and hence the outputs, are not guaranteed to be preserved when the order is swapped; in OMP, by contrast, RoPE never relates positions across chains (Section 3.2), so no such offset arises. In practice, order dependence can also arise after the encoder: the MINT representations we measured are invariant up to numerical error, but its prediction head concatenates the two embeddings in input order (Appendix E). For concatenation-based models, invariance is typically encouraged through data augmentation (Liu et al., 2025), which does not guarantee it, or imposed by averaging the predictions for both orders, which doubles the inference cost.

H Domain-specific prior art for the downstream tasks

Each of our downstream tasks has its own specialized literature, which we summarize here and contrast with the setting evaluated in the main text.

TCR–epitope binding. ProtLM.TCR (Essaghir et al., 2022) pre-trains a RoBERTa-style language model on TCR sequences and predicts binding with six layers that interleave self-attention, cross-attention and feed-forward blocks over separately embedded epitope and TCR sequences; it is the closest architectural precedent of OMP that we are aware of, applied to a single domain and trained end-to-end for it. LANTERN (Qi et al., 2026) encodes the TCR with ESM and the peptide as a SMILES string with a molecular language model, and aligns the two representations through multi-head cross-attention. LANTERN is evaluated on TCHard, the collection behind the TCR–epitope task of TDC-2 that we use, but its peptide encoding and its train–test splits differ from the "RN" splits of TDC-2, so its reported numbers are not directly comparable with ours.

Interface prediction. Pair-EGRET (Alam et al., 2024) predicts interaction sites with an edge-aggregated graph attention network over a k-nearest-neighbour graph of the native structure, combining protein language model features with cross-attention on a Siamese network. Because it consumes three-dimensional structure as input, it addresses a different regime from the sequence-only setting of the main text, where structure is used only to define the ground-truth labels.

What their published numbers show. LANTERN is evaluated on TCHard under four settings, of which the original random control with a hard epitope split is the closest to the "RN" set of TDC-2 that we use: the negatives come from TCHard itself, the inputs are the TCR beta chain and the peptide as in our setup, and peptides in the test set never appear in training. On that setting, Qi et al. (2026) report an AUROC of 0.570 for LANTERN and of 0.573, 0.562, 0.535, 0.531, 0.530, 0.521, 0.501 and 0.501 for Smiles-Bert, Imrex, NetTCR, TEINet, ChemBERTa, ERGO II, TITAN and DlpTcr, along with an AUC-PR of 0.400 for LANTERN. The protocols are close but not identical: they keep the original TCHard test set fixed and split the training set five ways, whereas TDC-2 publishes five train–test splits. We therefore quote these values as context rather than entering them in our table. For interface prediction no such transfer is possible: Pair-EGRET reports on DBD5, the Dockground unbound benchmark and a subset of MaSIF, and not on PINDER, so its numbers cannot be placed beside ours.

Why we do not reproduce them. Each of these models is trained for a single task, with its own inputs, splits and, in two cases, modalities that OMP does not receive. Our comparison isolates the pre-trained representation instead: the same frozen backbone and the same prediction head across three domains, against ESM-2 and MINT under identical conditions. Benchmarking specialized predictors against pair-level protein language models, each in its own regime, is a natural next step that we leave for future work.

I Supplementary figures and tables

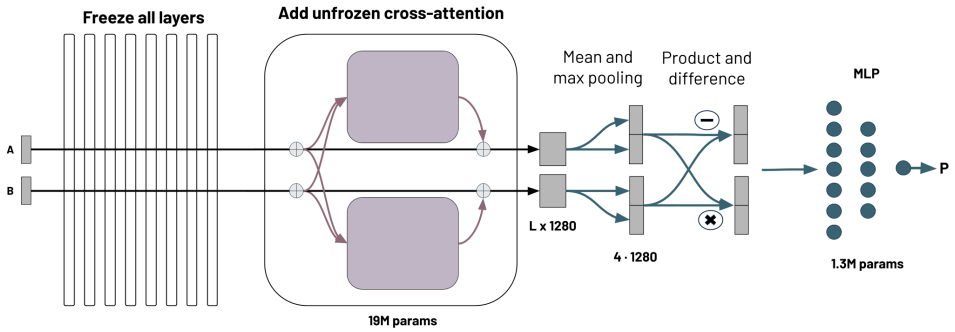


Figure 5: Architectural scheme of the Siamese order-invariant affinity module classification head attached to the frozen foundational encoder backbone.

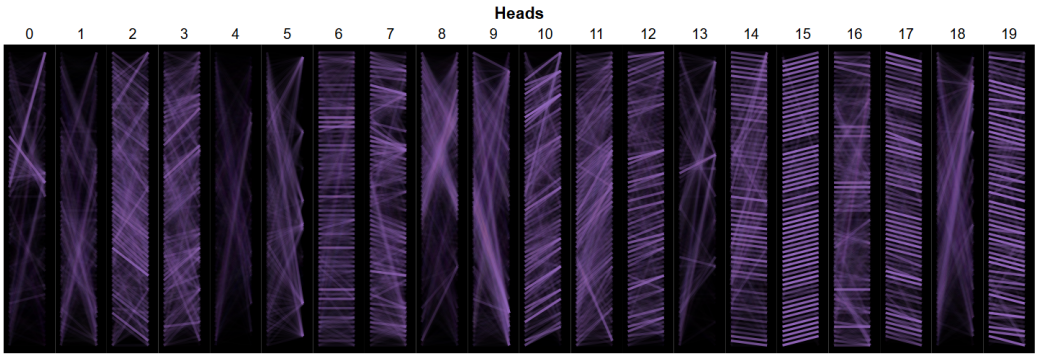


Figure 6: Functional specialization of the multi-head mechanism at layer 30 of OMP self-attention, mapped from amino acid index 105 to 168 for protein X as input.

Table 6: Self-attention specialization metrics for the 20 standard amino acids in OMP. The columns define: the background frequency of each residue within the testing set; the top specialized layer-head index; the absolute percentage of self-attention directed to that residue; the relative bias multiplier (top head attention divided by background frequency); and the ratio of top head attention to the global head mean.

Code	Abbrev	Name	Background (%)	Top Head	Top Self-Attn (%)	Relative Bias	Top/Mean
A	Ala	Alanine	8.9	30-18	24.9	2.8	3.1
C	Cys	Cysteine	1.3	3-17	22.2	17.1	10.7
D	Asp	Aspartic acid	5.6	4-9	35.9	6.4	6.5
E	Glu	Glutamic acid	6.3	9-8	34.1	5.4	5.7
F	Phe	Phenylalanine	3.8	28-14	26.4	6.9	5.8
G	Gly	Glycine	7.1	32-3	41.4	5.8	5.2
H	His	Histidine	2.2	0-12	9.7	4.4	3.4
I	Ile	Isoleucine	5.4	10-9	13.8	2.6	2.1
K	Lys	Lysine	5.0	30-5	35.9	7.2	6.8
L	Leu	Leucine	9.7	1-14	48.3	5.0	4.7
M	Met	Methionine	2.2	29-6	22.3	10.1	8.9
N	Asn	Asparagine	4.0	4-9	11.7	2.9	3.0
P	Pro	Proline	5.1	6-16	21.3	4.2	5.0
Q	Gln	Glutamine	3.9	9-13	15.3	3.9	4.6
R	Arg	Arginine	5.8	30-5	29.0	5.0	6.3
S	Ser	Serine	7.3	6-10	18.8	2.6	3.4
T	Thr	Threonine	5.7	6-10	12.0	2.1	2.5
V	Val	Valine	6.7	10-9	15.3	2.3	2.1
W	Trp	Tryptophan	1.2	28-14	15.3	12.8	9.0
Y	Tyr	Tyrosine	2.9	28-14	25.1	8.7	6.9

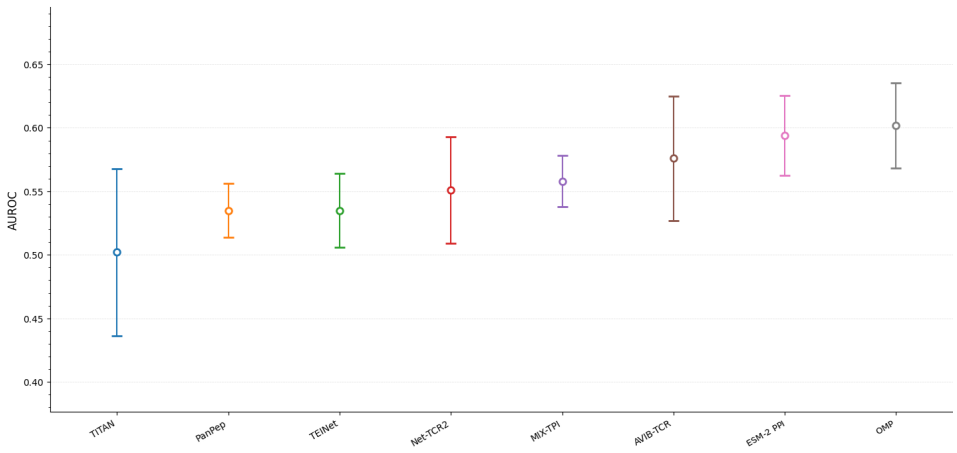


Figure 7: Error bar plot showing the mean AUROC in the TCR–epitope interaction prediction task, tracking standard deviation margins across the five test splits. Baseline metrics are reproduced directly from Velez-Arce et al. (2024).

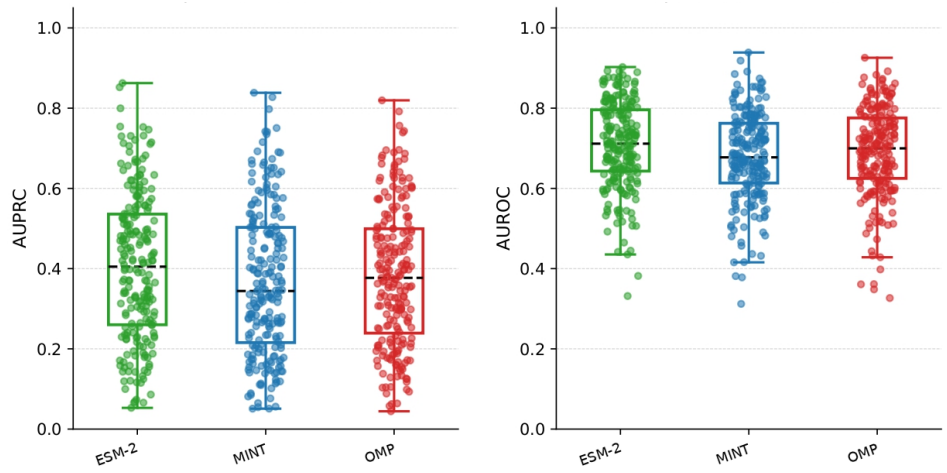
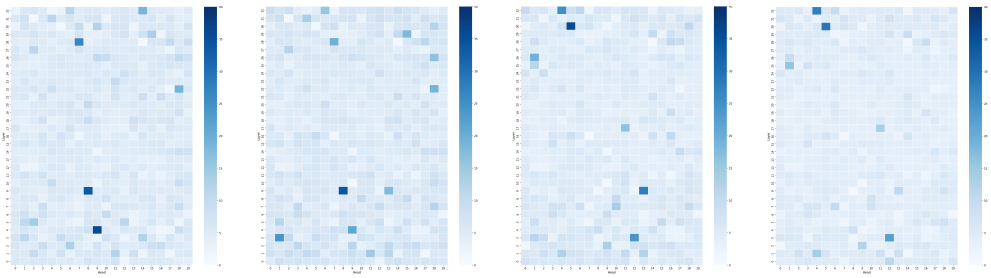


Figure 8: Box plots illustrating the distribution of per-sequence-pair AUPRC and AUROC within the cropped PINDER S test set ($n = 220$) for ESM-2, MINT, and OMP. These models achieve median AUPRC values of 0.4049, 0.3442, and 0.3770, and median AUROC values of 0.7115, 0.6774, and 0.6998, respectively. Statistical analysis indicates that ESM-2 achieves a higher AUROC than OMP (Wilcoxon p-value of 0.000697), while OMP outperforms MINT (Wilcoxon p-value of 0.008642).



(a) Activation for asp. (b) Activation for glu. (c) Activation for lys. (d) Activation for arg.

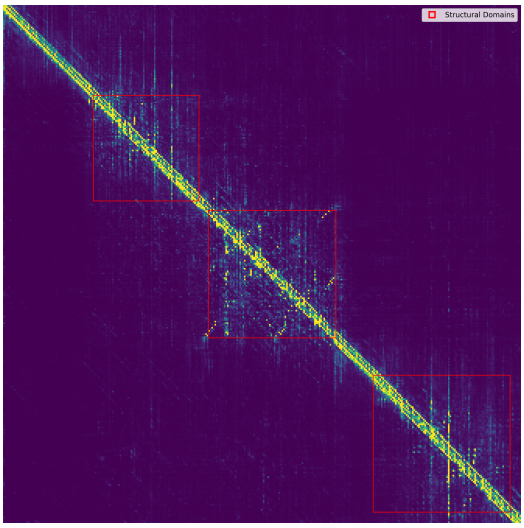
Figure 9: Self-attention distributions across layers and heads of OMP for **a** Aspartic acid, **b** Glutamic acid, **c** Lysine, and **d** Arginine.

Table 7: Pearson correlation coefficients between empirical self-attention similarity matrices and the theoretical BLOSUM62 substitution matrix across baseline architectures (Henikoff & Henikoff, 1992). Baseline metrics are reproduced from Vig et al. (2021), while ESM-2 and OMP were calculated empirically.

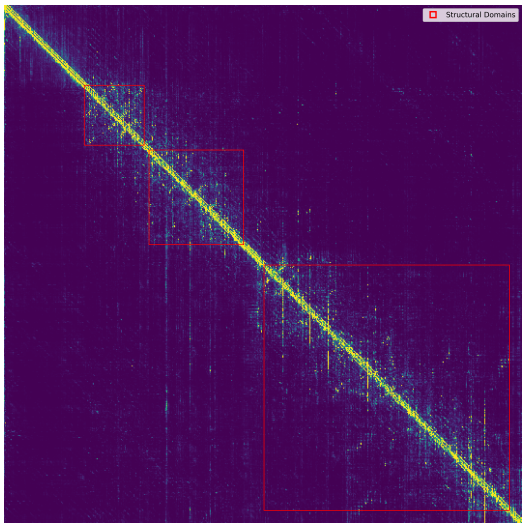
Model	Pearson Correlation
ProtAlbert	0.60
ProtBERT	0.68
ProtXLNet	0.71
TapeBERT	0.73
ProtBERT-BFD	0.75
ESM-2	0.78
OMP	0.79

Table 8: Amino acid sequences in the UniProt database (https://www.uniprot.org/) used for interpretability (Barrett et al., 2013; Roskoski, 2015). The identities of proteins X and Y are temporarily withheld to preserve author anonymity during double-blind review and will be disclosed after the review process.

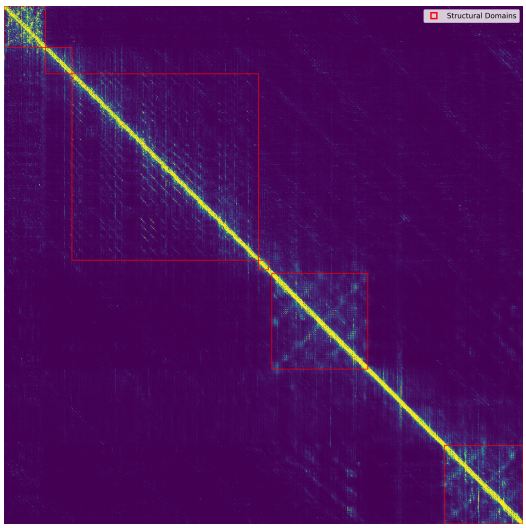
Protein	Sequence
BCAR1	MNHLNVLAKALYDNVAESPDELSFRKGDIMTVLEQDTQGLDGWWLCSLHGRQGIVPGNRLKILVGMYDKKPAGPG PGPPATPAQPQPGLHAPAPPASQYTPMLPNTYQPQPDSVYLVPTPSKAQQGLYQVPGPSPQFQSPPAKQTSTFSKQTPH HPFPSPATDLYQVPPGPGGPAQDIYQVPPSAGMGHDIYQVPPSMDTRSWEGTKPPAKVVVPTRVGQGYVYEAAQPEQ DEYDIPRHLLAPGPQDIYDVPPVRGLLPSQYGQEVYDTPPMAVKGPNGRDPLLEVYDVPPSVEKGLPPSNHHAVYDV PPSVSKDVPDGPLLREETYDVPPAFAKAKPFDPARTPLVLAAPPPDSPPAEDVYDVPPPAPDLYDVPPGLRRPGPGTLY DVPRERVLPPEVADGGVVDSGVYAVPPPAEREAPAEGKRLSASSTGSTRSSQSASSLEVAGPGREPLELEVAVEALAR LQQGVSATVAHLLDLAGSAGATGSWRSPSEPQEPLVQDLQAAVAAVQSAVHELLEFARSAVGNAAHTSDRALHAKL SRQLQKMEDVHQTLVAHGQALDAGRGGSGATLEDLDRLVACSRAVPEDAKQLASFLHGNASLLFRRTKATAPGPEG GGTLHPNPTDKTSSIQSRPLPSPPKFTSQDSPDGQYENSEGGWMEDYDYVHLQGKEEFEKTQKELLEKGSITRQGKSQ LELQQLKQFERLEQEVSRPIDHDLANWTPAQPLAPGRTGGLGPSDRQLLLFYLEQCEANLTTLTNAVDAFFTAVATN QPPKIFVAHSKFVILSAHKLVFIGDTLSRQAKAADVRSQVTHYSNLLCDLLRGIVATTKAAALQYPSPSAAQDMVERV KELGHSTQQFRRVLGQLAAA
SRC	MGSNKSKPKDASQRRRSLEPAENVHGAGGGAFPASQTPSKPASADGHRGPSAAFAPAAAEPKLFGGFNSSDTVTSPQ RAGPLAGGVTTFVALYDYESRTETDLSFKKGERLQIVNNTEGDWWLAHSLSTGQTGYIPSNYVAPSDSIQAEEWYFG KITRRESERLLLNAENPRGTFLVRESETTKGAYCLSVSDFDNAKGLNVKHYKIRKLDSGGFYITSRTQFNSLQQLVAY YSKHADGLCHRLTTVCPTSKPQTQGLAKDAWEIPRESLRLEVKLGQGCFGEVWMGTWNGTTRVAIKTLKPGTMSPE AFLQEAQVMKKLRHEKLVQLYAVVSEEPIYIVTEYMSKGSLLDFLKGETGKYLRLPQLVDMAAQIASGMAYVERMN YVHRDLRAANILVGENLVCKVADFGLARLIEDNEYTARQGAKFPIKWTAPEAALYGRFTIKSDVWSFGILLTELTTKG RVPYPGMVNREVLDQVERGYRMPCPPECPESLHDLMCQCWRKEPEERPTFEYLQAFLEDYFTSTEPQYQPGENL



(a) Self-attention activation at layer 30 for protein Y with domain overlay.

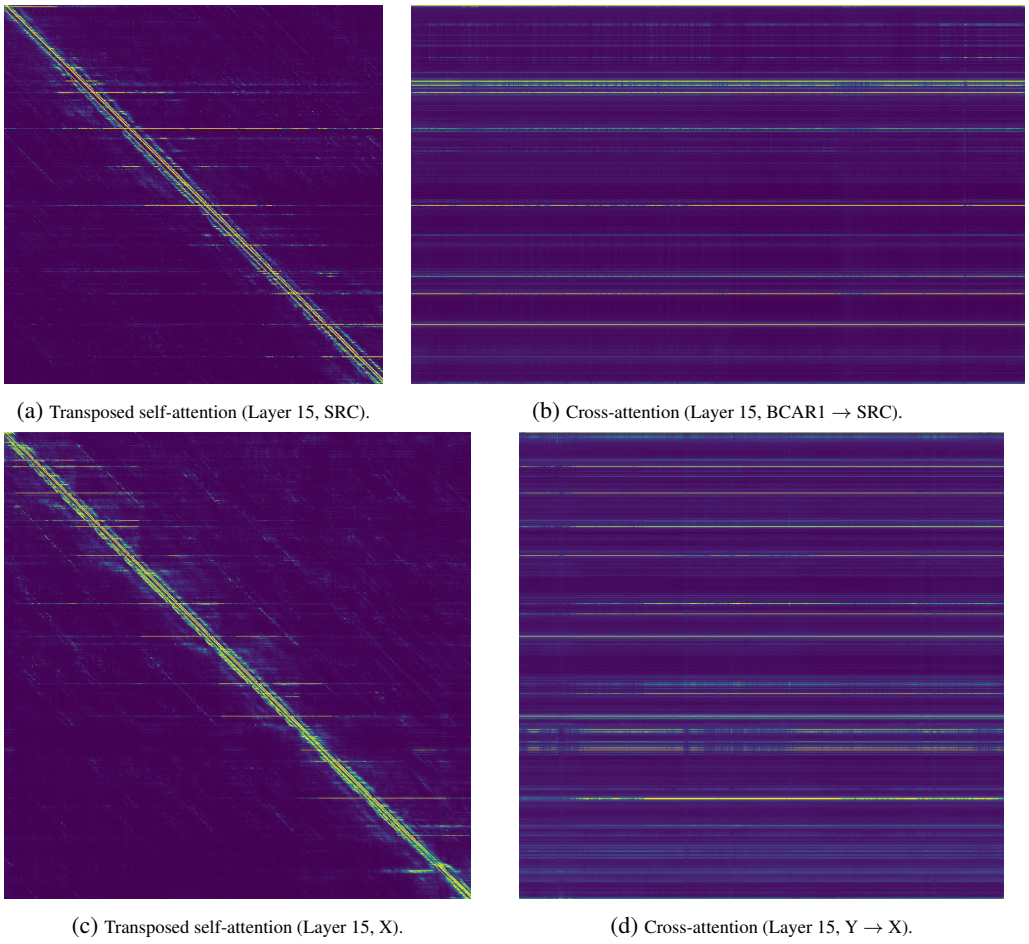


(b) Self-attention activation at layer 30 for SRC with domain overlay.



(c) Self-attention activation at layer 30 for BCAR1 with domain overlay.

Figure 10: Zero-shot structural interpretability of OMP using protein pairs. **a** Heatmap displaying raw self-attention activation at layer 30 for protein Y with domain overlay, utilizing the X–Y complex as input. **b** Heatmap displaying raw self-attention activation at layer 30 for SRC with domain overlay (SH3, SH2, SH1 catalytic), utilizing the SRC–BCAR1 complex as input (Roskoski, 2015; Schuh et al., 2010; Nakamoto et al., 1996). **c** Heatmap displaying raw self-attention activation at layer 30 for BCAR1 (p130Cas) with domain overlay (SH3, Proline-rich, Substrate, Serine-rich, FAT-like, FAT-like), utilizing the SRC–BCAR1 complex as input (Barrett et al., 2013; Briknarová et al., 2005; Mace et al., 2011; Schuh et al., 2010; Nakamoto et al., 1996).



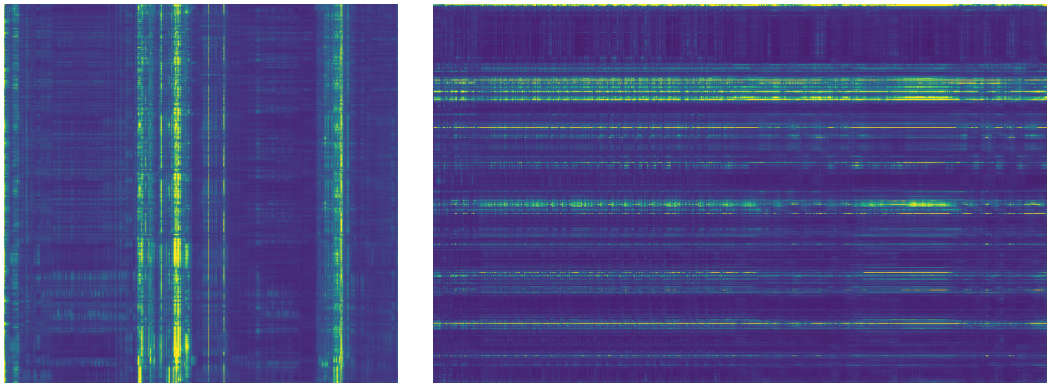
(a) Transposed self-attention (Layer 15, SRC).

(b) Cross-attention (Layer 15, BCAR1 → SRC).

(c) Transposed self-attention (Layer 15, X).

(d) Cross-attention (Layer 15, Y → X).

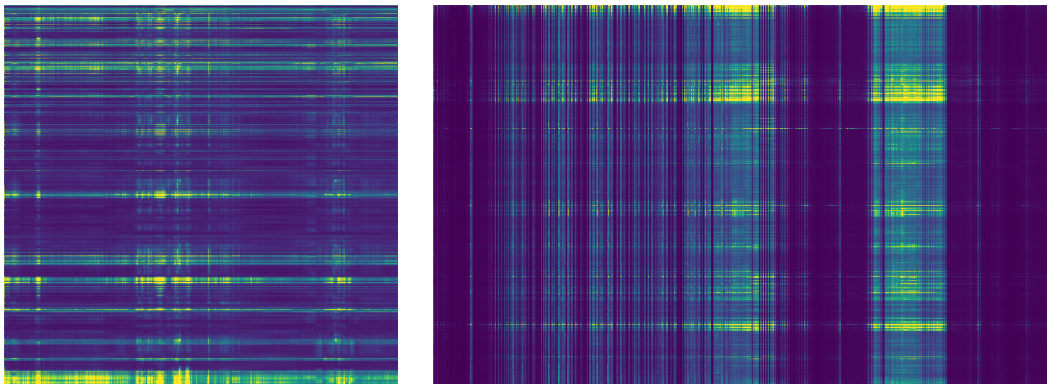
Figure 11: Zero-shot comparative self- versus cross-attention interpretability profiles at layer 15 of OMP. **a** Heatmap displaying transposed raw self-attention activation at layer 15 for the SRC sequence. **b** Heatmap showing cross-attention activation at layer 15 for the BCAR1 sequence (x-axis) querying the SRC sequence (y-axis). **c** Heatmap displaying transposed raw self-attention activation at layer 15 for the protein X sequence. **d** Heatmap showing cross-attention activation at layer 15 for the protein Y sequence (x-axis) querying the protein X sequence (y-axis).



(a) Cross-attention (Layer 20, X → Y).

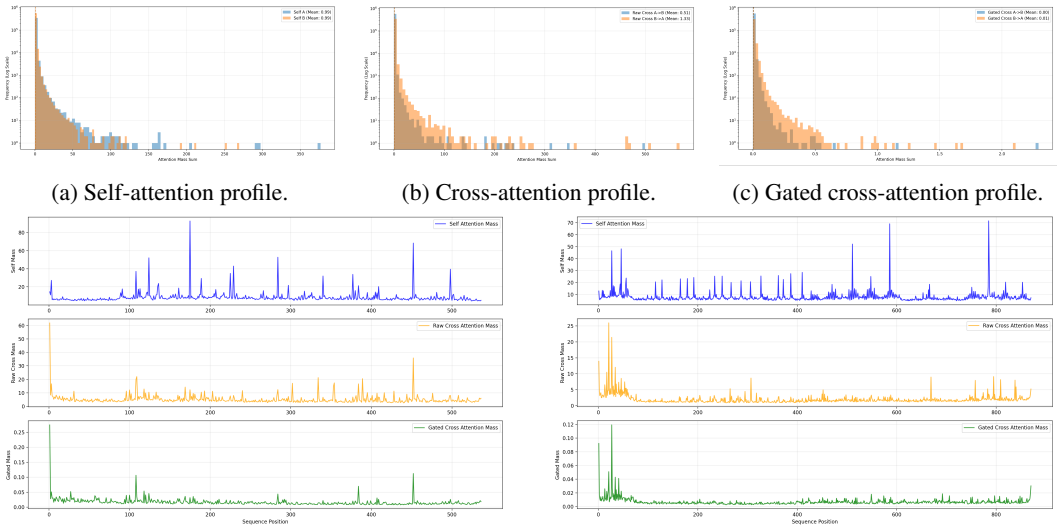
(b) Cross-attention (Layer 20, BCAR1 → SRC).

Figure 12: Zero-shot cross-attention interpretability profiles at layer 20 of OMP. **a** Heatmap showing cross-attention weights at layer 20 for the protein X sequence (y-axis) querying the protein Y sequence (x-axis). **b** Heatmap displaying cross-attention activation in layer 20 for BCAR1 (x-axis) querying SRC (y-axis).



(a) Gated cross-attention (Layer 21, X → Y). (b) Gated cross-attention (Layer 21, BCAR1 → SRC).

Figure 13: Zero-shot gated cross-attention interpretability profiles of OMP. **a** Heatmap showing gated cross-attention weights at layer 21 for the protein X sequence (y-axis) querying the protein Y sequence (x-axis). **b** Heatmap displaying gated cross-attention activation in layer 21 for BCAR1 (x-axis) querying SRC (y-axis).



(a) Self-attention profile. (b) Cross-attention profile. (c) Gated cross-attention profile.

(d) Positional alignment for SRC. (e) Positional alignment for BCAR1.

Figure 14: Comprehensive attention sink and positional bias profiling within the SRC–BCAR1 complex case study. **a–c** Log-scale frequency histograms mapping token-level attention densities to isolate sink signatures across self-, cross-, and gated cross-attention modalities. **d** Positional attention mass tracking across the linear residue index of Chain A (SRC) to isolate terminal boundary localization biases. **e** Positional attention mass tracking across the linear residue index of Chain B (BCAR1).

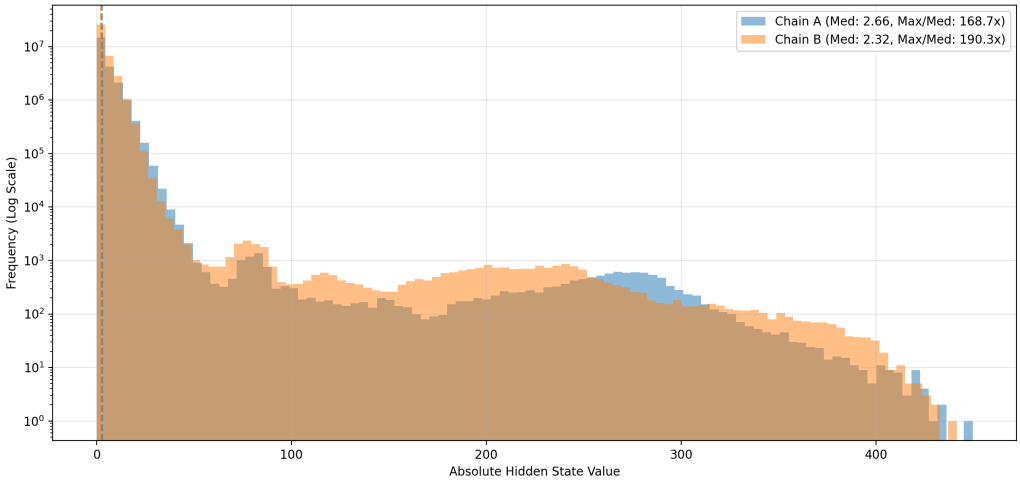


Figure 15: Log-scale frequency distribution of empirical activation values extracted from the transformer hidden states during a forward pass of the SRC–BCAR1 complex, where Chain A denotes the SRC sequence and Chain B denotes the BCAR1 sequence.