RETHINKING PROTEIN SIMILARITY: ONE REPRESENTATION, MANY RELATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein retrieval depends on the biological question: finding proteins with a similar fold and finding proteins that bind the same ligand can require different rankings of the same candidates. A fixed similarity score cannot adjust its ranking to the requested criterion, limiting retrieval tailored to different biological questions. We introduce ReaPO (Relation-adaptive Protein Operators), which uses a natural-language instruction and a few directed example comparisons to adapt the similarity rule over frozen protein language model representations. Its signed variant, ReaPO-Signed, assigns relation-dependent positive or negative weights to individual feature directions. We evaluate the approach on crossed ranking, which requires opposite correct rankings for the same query and candidates under two relations, and on ordinary protein retrieval using internal and external datasets. With both relation families withheld from readout training, ReaPO-Signed achieves 0.548 joint accuracy, compared with 0.514 for a conditional positive-semidefinite scorer augmented with a learned global signed scalar that can reverse the whole ranking. When trained separately using only a target-ranking objective, ReaPO-Signed improves nDCG@10 over matched condition-independent signed readouts by 0.0236 on Ordinary20 and 0.0309 on external CATH–BioLiP2 retrieval in the Remote setting. These results support adapting comparison rules to biological relations while reusing shared frozen protein representations.

1 INTRODUCTION

Protein retrieval seeks proteins with a desired structural or functional relationship to a query. The same query and candidates may require different rankings for shared structure, molecular function, ligand context or cellular role. A condition-independent score cannot adjust its ranking to the requested criterion, limiting retrieval tailored to different biological questions. Sequence similarity alone, for example, can miss functional substitutes in cross-species complementation (Kachroo et al., 2015; Shaw et al., 2025).

Protein language models (PLMs) encode information about multiple biological properties, including structure and function (Rives et al., 2021; Lin et al., 2023; Elnaggar et al., 2022; Vig et al., 2021). A shared representation does not imply a unique notion of similarity: a fixed scorer applies one comparison rule to features that may support several relations. Conditional Similarity Networks and GeneCIS study criterion-dependent comparison in vision (Veit et al., 2017; Vaze et al., 2023). We ask whether instructions and a few directed examples can specify how frozen protein features should be compared.

We introduce ReaPO (Relation-adaptive Protein Operators), which compares frozen protein features using a natural-language instruction and a few examples. Each example specifies which protein is preferred relative to an anchor. Given an instruction and a support set, ReaPO generates one operator and reuses it across protein pairs. Its signed variant, ReaPO-Signed, assigns positive or negative weights to individual feature directions. Users with reliable annotations or expert knowledge can assemble support from known proteins to express the desired comparison criterion.

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.

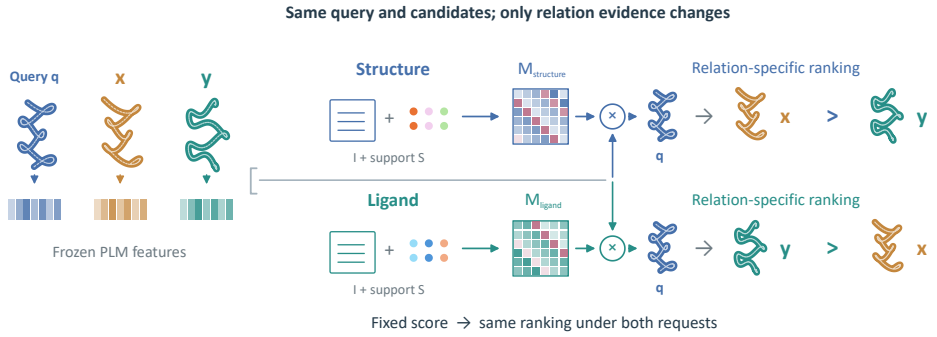


Figure 1: **Same proteins, different comparison rules.** Only the instruction and directed support change; query q, candidates x, y and their frozen feature bank remain fixed. Protein ribbons, operator entries and rankings are schematic.

We use two complementary evaluations. *Crossed ranking* tests relation responsiveness by fixing the query and candidates and requiring opposite correct rankings under two relations (Figure 1). Both relation families are withheld throughout readout training. Ordinary retrieval tests whether conditioning remains useful when candidates need not satisfy opposing preferences.

Under strict held-out-family crossed ranking, ReaPO-Signed reaches **0.548 joint accuracy**, versus **0.514** for conditional positive-semidefinite (PSD) scoring with a learned global signed scalar (Table 1), supporting gains beyond whole-score reversal under matched training. With separate target-ranking-only training, it reaches **0.596 nDCG@10** on Ordinary20, improving over its matched condition-independent signed readout by **0.0236** (Table 3). The gain is **0.0309 nDCG@10** on external CATH–BioLiP2 in the Remote retrieval setting, without conflict-selected candidates (Table 4).

Our contributions are:

- **Relation-conditioned comparison over frozen representations.** ReaPO infers comparison rules from instructions and directed examples; ReaPO-Signed implements them with signed diagonal-plus-low-rank operators and bounded, normalized scores.
- **Controlled evaluation of relation switching.** Crossed ranking fixes the query and candidates while requiring opposite rankings under two relations, directly testing responsiveness to relation evidence under held-out relation families.
- **Retrieval beyond conflict-selected examples.** Separately trained target-ranking models test whether conditioning remains beneficial in ordinary gallery retrieval, internally and on external CATH–BioLiP2 evaluation.

2 Related Work

Protein representations and geometry. TAPE evaluates protein representations across tasks and splits (Rao et al., 2019). Retrieval-focused work such as Protriever couples learned homology retrieval to a downstream objective (Weitzman et al., 2025), while studies of PLM geometry examine variation across depth (Beshkov & Malthe-Sorenssen, 2026). Our focus is relation-conditioned scoring over a frozen feature bank.

Conditional similarity and few-shot adaptation. Conditional Similarity Networks learn visual similarity subspaces (Veit et al., 2017), and GeneCIS evaluates text-specified criteria (Vaze et al., 2023). Few-shot methods form prototypes or adapt support representations (Snell et al., 2017; Ye et al., 2020). In ReaPO, set attention encodes directed support comparisons (Lee et al., 2019), and a hypernetwork generates the coefficients of an operator shared across query–candidate pairs (Ha et al., 2017).

Language and aspect-aware protein retrieval. ProtST aligns sequence and biomedical text (Xu et al., 2023); ProTrek jointly represents sequence, structure and text (Su et al., 2026); Protein-Vec combines aspect-specific embeddings (Hamamsy et al., 2023). InstructProtein studies instruction-

tuned protein–text generation (Wang et al., 2024). We compare frozen sequence scores and instruction adapters under a common ranking protocol, with input and coverage differences specified in Appendix B.

3 Relation-Conditioned Readout with Instructions and Examples

3.1 Task, representations and directed evidence

Task. Given a query protein q, candidate proteins x, and a condition $\mathcal{C} = (I, \mathcal{S})$, the task is to rank candidates according to the specified biological relation. Here I is a natural-language instruction and $\mathcal{S} = \{(a_k, p_k, n_k)\}_{k=1}^{K}$ contains K support triples. In triple k, p_k is preferred to n_k relative to anchor a_k. The conditioner receives no relation identifier, family index or query–candidate relevance label. Sequence components separate support proteins from query and candidate roles (Section 4).

Model overview. Instruction and support generate an operator shared across query–candidate pairs under the same condition, defining $s_{\mathcal{C}}(q, x)$ (Figure 2). The protein and text backbones are frozen; the support encoder, text adapter, fusion, coefficient heads and shared projection basis are trained. No parameters are updated at inference.

Representations. We mean-pool residue states, excluding special tokens, at ESM2-650M layers $\mathcal{L} = (0, 6, 12, 18, 24, 30, 33)$, sampling depth every six layers and retaining the final layer. Per-layer normalization and concatenation give $h(x) \in \mathbb{R}^D$, with $D = 8960$. Protein features and mean-pooled, normalized PubMedBERT instruction features (Gu et al., 2021) are cached.

Each support triple is encoded as

$$e_k = h(a_k) \odot \big[h(p_k) - h(n_k)\big]. \tag{1}$$

The anchor makes the contrast reference-dependent; swapping p_k and n_k negates e_k. For diagonal weights d, the unnormalized bilinear margin $h(a_k)^\top \operatorname{diag}(d)[h(p_k) - h(n_k)] = d^\top e_k$ motivates this encoding. Section 4.3 tests the anchor and the positive–negative contrast.

3.2 Inducing a relation-specific similarity operator

The support encoder maps contrasts to tokens, applies self-attention without positional encodings and mean-pools them into $v \in \mathbb{R}^{64}$. A learned adapter maps the frozen instruction embedding to $u \in \mathbb{R}^{64}$. A two-layer MLP fuses them:

$$c_{\mathcal{C}} = f_\theta([v; u; v \odot u]). \tag{2}$$

The resulting condition vector is invariant to support order in exact arithmetic. The condition generates coefficients for

$$M_{\mathcal{C}} = \operatorname{diag}(d_{\mathcal{C}}) + B^\top \operatorname{diag}(\lambda_{\mathcal{C}}) B, \tag{3}$$

where $d_{\mathcal{C}} \in \mathbb{R}^D$, $\lambda_{\mathcal{C}} \in \mathbb{R}^{64}$ and $B \in \mathbb{R}^{64 \times D}$, with rows $b_r^\top$, is learned across training conditions. The diagonal term reweights coordinates; each low-rank term $\lambda_{\mathcal{C},r}(b_r^\top h(q))(b_r^\top h(x))$ weights a shared direction b_r. Conditioning generates coefficients, not a dense matrix.

Two heads generate $d_{\mathcal{C}}$ and $\lambda_{\mathcal{C}}$ using $\tanh$ followed by per-head mean-absolute normalization (Appendix A). This controls coefficient scale; individual coefficients can exceed one.

The matched PSD control takes absolute values of both heads, constraining all coefficients to be nonnegative. Global-sign PSD additionally allows one condition-dependent scalar to reverse the whole score; Signed allows different coefficient signs within one operator. Coefficient signs should not be interpreted as eigenvalue signs or biological mechanisms. PSD can also reverse rankings across conditions; we test the empirical value of coefficient-specific signs (Appendix A.4).

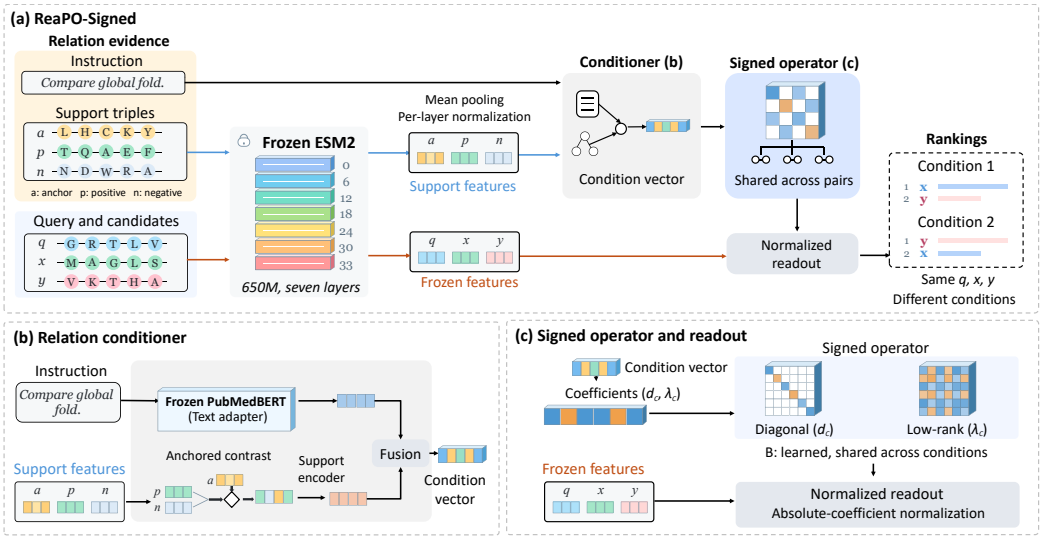


Figure 2: **ReaPO-Signed architecture.** (a) Relation-conditioned ranking. (b) Relation conditioner with frozen PubMedBERT and a trainable text adapter. (c) Condition-generated coefficients weight a basis learned across conditions. Linked circles denote protein pairs sharing one operator under the same instruction and support. Rankings show different conditions for the same query and candidates. All diagrams are schematic.

3.3 Bounded scoring and retrieval

Indefinite $M_{\mathcal{C}}$ can give negative self-quadratic forms. Normalization therefore uses the positive-semidefinite operator

$$H_{\mathcal{C}} = \operatorname{diag}(|d_{\mathcal{C}}|) + B^{\top} \operatorname{diag}(|\lambda_{\mathcal{C}}|) B, \quad n_{\mathcal{C}}(x) = \sqrt{\max\{h(x)^{\top} H_{\mathcal{C}} h(x), 10^{-8}\}},$$
$$s_{\mathcal{C}}(q,x) = \kappa \frac{h(q)^{\top} M_{\mathcal{C}} h(x)}{n_{\mathcal{C}}(q) n_{\mathcal{C}}(x)}, \quad \kappa = \min(e^{\beta}, 20). \tag{4}$$

Here β is learned; $H_{\mathcal{C}}$ generally differs from the spectral absolute value $|M_{\mathcal{C}}|$. Cauchy–Schwarz in the diagonal and projected coordinates gives

$$\left|h(q)^{\top} M_{\mathcal{C}} h(x)\right| \le \sqrt{h(q)^{\top} H_{\mathcal{C}} h(q)} \sqrt{h(x)^{\top} H_{\mathcal{C}} h(x)}. \tag{5}$$

Hence $|s_{\mathcal{C}}(q,x)| \le \kappa$, including zero vectors. The score is symmetric. PSD has $M_{\mathcal{C}} = H_{\mathcal{C}} \succeq 0$ and defines a kernel for each fixed condition; Signed permits indefinite similarities (Oglic & Gaertner, 2018). Both variants define similarity scores; a distance requires a separate definition.

For each condition, $s_{\mathcal{C}}(q,x) = \langle \phi^{q}_{\mathcal{C}}(q), \phi^{c}_{\mathcal{C}}(x) \rangle$, with both factors in $\mathbb{R}^{D+64}$ (Appendix A.1). This dot-product form allows candidate factors to be cached across queries. Changing instruction or support rebuilds the factors while reusing frozen protein features.

3.4 Learning the readout

Separate readouts are trained for the two tasks. Crossed ranking uses switching and condition-preference losses; ordinary retrieval uses only family-balanced $\mathcal{L}_{\text{target}}$ ("target-only"), with instruction and support retained in conditional readouts. Fixed-Signed and Fixed-PSD zero both condition inputs before fusion throughout training and inference; fusion biases and geometry remain learnable.

Let $g_{\mathcal{C}}(q,p,n) = \tau[s_{\mathcal{C}}(q,p) - s_{\mathcal{C}}(q,n)]$, with learned positive training scale τ, and $\ell_{\delta}(z) = \operatorname{softplus}(\delta - z)$. Target ranking and preference for correct over corrupted conditions use

$$\mathcal{L}_{\text{target}} = \ell_{.2}(g_{\mathcal{C}}), \qquad \mathcal{L}_{\text{condition}} = \ell_{\delta}(g_{\mathcal{C}} - g_{\widetilde{\mathcal{C}}}), \tag{6}$$

where $\widetilde{\mathcal{C}}$ changes the instruction or support, or reverses support preferences, and δ is stage-specific. For a switch, $\mathcal{C}_x$ prefers x and $\mathcal{C}_y$ prefers y for the same query. Writing $g_x = g_{\mathcal{C}_x}(q,x,y)$ and

$g_y = g_{\mathcal{C}_y}(q, x, y)$, the core objective is

$$\mathcal{L}_{\text{switch,core}} = \tfrac{1}{2}[\ell_{.2}(g_x) + \ell_{.2}(-g_y)] + \ell_{.2}(g_x - g_y). \tag{7}$$

This encourages $g_x > 0$ and $g_y < 0$ and separates their margins. Corrupted-condition preferences compare target-oriented margins: g_x for $\mathcal{C}_x$ and $-g_y$ for $\mathcal{C}_y$, applying the same orientation to each corrupted margin.

Crossed training has three stages. **Annotation training** learns target ranking, condition preference, instruction–support alignment and condition contrast from relation-labelled triples. **Broad switch training** learns opposite rankings for the same proteins.

Switch refinement adds preference over *annotation-mismatched support* (code name: "wrong-mechanism"): equally sized, role-disjoint donors from another same-relation annotation group, holding instruction, query and candidates fixed. Matched donors share a target query–positive annotation; replacements use a label absent from the query. This auxiliary loss assumes annotation matching supplies more relevant context, without treating other same-relation examples as mislabeled (Appendix A.3).

Crossed training alternates relation-specific instructions with a generic prompt asking the model to infer the criterion from support. Both held-out families are excluded; Signed, PSD and Unrestricted use matched supervision and schedules. Final scheduled checkpoints are used without evaluation-label selection. Appendix A.3 gives objectives, corruptions and budgets; Appendix A.4 defines controls.

4 Experiments

4.1 Evaluation protocols

Crossed ranking and ablations use three-stage models to test whether rankings follow the requested relation. Ordinary20 and external retrieval test whether conditioning helps when candidate selection does not require opposite rankings under two relations. They use separately trained conditional and fixed readouts, with only the target-ranking loss.

Held-out testing. The internal crossed-ranking and Ordinary20 test split was excluded from readout training, method development, hyperparameter tuning and checkpoint selection.

Data and separation. Both internal evaluations use Core20, our dataset of 20 annotation-based relations grouped into ten relation families for training holdout (e.g., bound-ligand and cofactor identity form the ligand-context family). Sequence components combine MMseqs2 clusters at 30% identity and 80% bidirectional coverage with exact-sequence matches in Swiss-Prot 2026_02 (Steinegger & Söding, 2017). Support, query and candidate proteins occupy distinct components within each crossed record, although proteins may recur across records. Readout training and crossed evaluation share no proteins or components.

Crossed-set construction. For a relation pair (r_a, r_b), each query has annotations for both. Candidate x shares a query label under r_a and satisfies the negative criterion under r_b; y satisfies the converse. Negatives require nonempty, disjoint label sets; missing or ambiguous annotations are excluded. Eligibility depends on annotations and enough eligible support triples, without model scores; matching, sampling and per-stage overlap rules appear in Appendix B.1.

Strict crossed ranking. For fixed query q and candidates x, y, $\mathcal{C}_x$ requires x above y and $\mathcal{C}_y$ the converse. Strict leave-two-family-out (L2FO) evaluation uses 4,282 records from 43 cross-family relation pairs across all 19 eligible family-pair folds and three training seeds. Both target relation families are excluded from every readout training stage. Evaluation uses each held-out relation's instruction and four component-disjoint support triples from the same relation pair. Unlike during switch refinement, the shared labels of support anchor–positive and target query–positive pairs cannot overlap (Appendix B.1). Both margins use the same candidate order:

$$\begin{aligned} m_x &= s_{\mathcal{C}_x}(q, x) - s_{\mathcal{C}_x}(q, y), \qquad m_y = s_{\mathcal{C}_y}(q, x) - s_{\mathcal{C}_y}(q, y), \\ \text{Joint} &= \mathbf{1}[m_x > 0 \ \wedge \ m_y < 0]. \end{aligned} \tag{8}$$

Joint requires both correct rankings; ties fail. Directional accuracy averages the two correctness indicators. Order, $\mathbf{1}[m_x > m_y]$, tests whether the margin favoring x is larger under $\mathcal{C}_x$ than $\mathcal{C}_y$;

both rankings need not be correct. P_I and P_S compare target-oriented margins under correct and mismatched instructions and supports, respectively. Condition-independent scores have Joint zero.

Ordinary retrieval. Ordinary20 tests ordinary retrieval using Core20's relations and held-out proteins, without opposing-ranking requirements or model-score selection. Each relation has 64 queries and 32 candidates per query (1,280 queries total), with at least one positive and one negative per gallery.

Four support triples per seed are reserved first; their sequence components across seeds are excluded from query/candidate sampling. For each query, nDCG@10, Hit@1, MRR and AP are averaged over seeds and checkpoints trained without its target relation family, then over queries within relations, weighting relations equally (Appendix A.4).

Controls. Support-selected sign preserves or reverses fixed cosine rankings according to support examples. Learned global sign multiplies conditional PSD by one $I + S$-dependent scalar in $(-1, 1)$, rescaling or reversing the score without coefficient-specific signs. Fixed-Signed and Fixed-PSD learn one operator per checkpoint, shared across conditions; conditional versions generate it from $I + S$. Unrestricted uses the same features, instruction, support and supervision in a neural query–candidate scorer without a bilinear constraint (Appendix A). Appendix B.4 details the baselines.

Uncertainty. Crossed estimates average seeds within equally weighted relation pairs. Relation-node intervals account for dependence between pairs sharing a relation; family-node intervals check sensitivity to shared relation families. Ordinary20 uses paired family bootstraps with equal relation weights. Queries are not independent replicates. Appendix B defines two-sided intervals and corrections, including Ordinary20's two-contrast correction.

4.2 Relation switching on held-out families

Table 1: **Strict crossed ranking.** Means over 43 equally weighted relation pairs and 19 held-out-family folds; trained scorers average three seeds. The global-sign control multiplies conditional PSD by one continuous condition-dependent scalar; Signed allows coefficient-specific signs. Instruction/support preference metrics: Table 11.

Method	Joint	Dir.	Order
Cosine baselines with support-selected signs			
Final-layer cosine + support sign	0.305	0.530	0.305
Seven-layer cosine + support sign	0.458	0.616	0.458
Learned conditional scorers			
Unrestricted	0.477	0.680	0.730
ReaPO-PSD	0.130	0.543	0.727
ReaPO-PSD + global sign	0.514	0.695	0.734
ReaPO-Signed	**0.548**	**0.718**	**0.747**

ReaPO-Signed reaches 0.548 Joint across 43 relation pairs and 19 strict-L2FO folds, versus 0.514 for conditional PSD with a learned global signed scalar (Table 1). The gain is **0.034 (relation-node 95% CI** $[0.016, 0.053]$**)**, with family-node interval $[0.023, 0.046]$. This control adapts PSD geometry and can reverse entire rankings, so the remaining gain supports coefficient-specific signs under matched training.

Conditional PSD reaches Order 0.727 but Joint 0.130: preferences often shift in the required direction without making both rankings correct. A learned global signed scalar narrows the Joint-accuracy gap between PSD and Signed. Seven-layer cosine with a support-selected sign reaches 0.458, below Signed by 0.090 (corrected CI $[0.034, 0.146]$).

Signed improves over PSD with global sign in 37 of 43 relation pairs and 17 of 19 family pairs. Its mean gain remains positive with equal family-pair weights and after excluding evaluation pairs involving each specified family or family combination, without retraining (Appendix C.2).

Unrestricted reaches 0.477 Joint with the same inputs. Signed's advantage is 0.071 $[-0.006, 0.148]$: the lower bound exceeds the predeclared -0.02 noninferiority margin, but the interval includes zero and does not establish superiority (Appendix B).

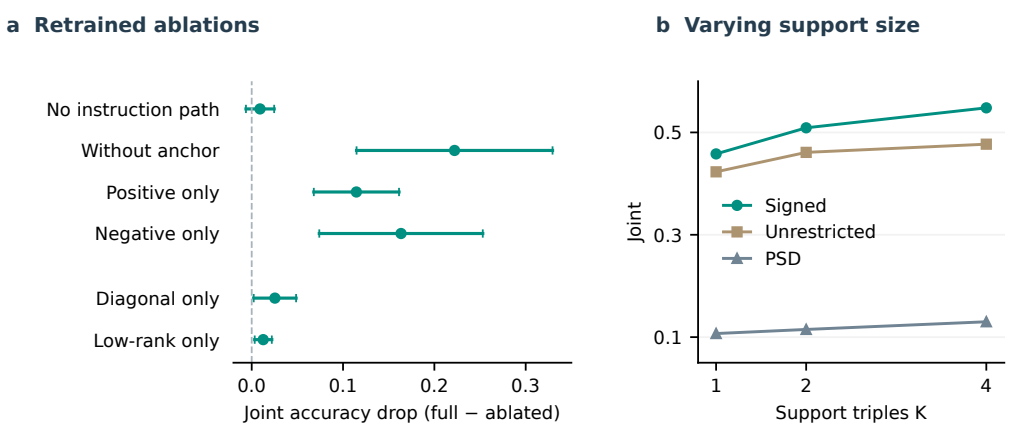


Figure 3: **Ablations and support-set size.** Left: Joint accuracy differences (complete model minus each retrained ablation), with six-contrast Bonferroni 99.1667% relation-node intervals; the instruction-path interval includes zero. Right: Joint for $K = 1, 2, 4$ support triples with the same trained models. All results use switching-trained readouts. Detailed metrics, family-node sensitivities and corrected support contrasts are in Appendix C.3.

4.3 What supports relation adaptation?

Table 2: **Pretraining and encoder controls.** Signed readouts trained separately on each frozen feature set; strict L2FO, three seeds. The first group tests pretrained information; the second uses other encoders with separately fitted readouts.

Feature source	Joint	Dir.	Order
Pretraining and information controls			
Pretrained ESM2-650M	0.548	0.718	0.747
Random ESM2-650M	0.260	0.513	0.518
Sequence CountSketch	0.260	0.529	0.536
Other pretrained encoders			
ProtT5-XL	0.547	0.714	0.743
ProtBERT	0.527	0.702	0.738

We next test the contribution of pretrained protein features. Replacing pretrained ESM2 with a randomly initialized ESM2 lowers Joint from 0.548 to 0.260 (Table 2); the paired difference is 0.288, with relation-node 95% CI $[0.184, 0.392]$. Compressed k-mer frequencies and sequence length (CountSketch) also reach 0.260. Both controls retain trained readouts, supporting the value of pretraining.

With separately trained Signed readouts, ProtT5 reaches 0.547 and ProtBERT 0.527, extending the result across three encoders. Appendix B.3 specifies all feature configurations.

Directed evidence and operator components. Figure 3 compares retrained ablations (left) and varying support size in the same trained models (right). Both use three-stage switching training.

Removing the anchor lowers Joint from 0.548 to 0.326; positive-only and negative-only support reach 0.434 and 0.385 (Table 17). These drops support using the anchor with both positive and negative examples. Diagonal-only and low-rank-only variants reach 0.523 and 0.536. Differences between the complete model and each variant have positive Bonferroni-adjusted relation-node and family-node intervals (Table 18).

Removing the instruction path yields 0.539 Joint. The 0.0091 advantage of the complete model is inconclusive under the primary adjusted relation-node interval, although the family-node sensitivity interval is positive. Language-related losses and mismatched-condition examples remain, so this tests the scoring path rather than all language supervision (Appendix C.3).

With the same trained models, increasing support from one to four triples raises Signed Joint from 0.458 to 0.548 (Table 19); no retraining is needed.

4.4 Ordinary retrieval performance and conditioning

With Signed and Fixed-Signed trained using only the target-ranking loss, ReaPO-Signed reaches 0.5959 nDCG@10 and exceeds separately trained Fixed-Signed by **0.0236** (Bonferroni-adjusted family-bootstrap CI $[0.0049, 0.0411]$; Tables 3 and 14). Signed, Fixed-Signed, PSD and Fixed-PSD share training records, initialization, all 19 folds, three seeds and the 12-epoch training budget.

Table 3: **Ordinary20 retrieval.** 20 equally weighted relations; 1,280 queries with 32 candidates each. I: instruction; S: four support triples. Operators use target-only training; adapters reuse three-stage checkpoints, using sequences and instructions without support (Appendix B.4).

Method	Inputs	nDCG@10	Hit@1	MRR
ESM2 seven-layer cosine	Seq.	0.472	0.398	0.529
ProtST official cosine	Seq.	0.554	0.437	0.585
ProTrek direct sequence	Seq.	0.466	0.379	0.520
ProtST adapter	Seq.+I	0.508	0.438	0.570
ProTrek adapter	Seq.+I	0.512	0.427	0.560
Fixed-PSD	Seq.	0.585	0.520	0.637
ReaPO-PSD	Seq.+I+S	**0.599**	**0.525**	**0.644**
Fixed-Signed	Seq.	0.572	0.500	0.622
ReaPO-Signed	Seq.+I+S	0.596	0.516	0.637

Scores use three decimals; bold marks column maxima. The two conditioning CIs target simultaneous 95% coverage using Bonferroni correction (Table 14). Other metrics are descriptive. Matched operator settings: Appendix A.4; AP and additional baselines: Table 13.



Figure 4: **Ordinary retrieval and conditioning.** Left: family nDCG@10; parentheses count relations. Right: descriptive Signed-minus-Fixed-Signed gains, including declines. Fixed-S: Fixed-Signed; PSD: ReaPO-PSD. The overall row averages 20 relations equally. Bold marks maxima; outlines mark Signed. Full results, including Fixed-PSD: Table 3 and Appendix C.1.

Gains occur in 17 of 20 relations, eight of ten families and all three seeds; the family differences remain negative for enzyme class and molecular function. Conditioning also improves PSD by 0.014 (Table 14); conditional PSD has a numerically higher mean than Signed, 0.599 versus 0.596.

For context, seven-layer cosine reaches 0.472 and ProtST sequence cosine 0.554 under the same evaluation protocol; these baselines differ in pretraining and inputs.

Family means vary (Figure 4): Signed exceeds ProtST for ligand context (0.474 versus 0.437), while ProtST is higher for molecular function (0.524 versus 0.469). Conditioning gains are largest in membrane topology and global structure. These family comparisons are descriptive.

4.5 External retrieval and transfer boundaries

Table 4: **External CATH–BioLiP2 retrieval.** nDCG@10 averaged equally over structure and ligand retrieval. Conditional and fixed operator readouts use target-only training. Paired 95% bootstrap intervals for Signed minus Fixed-Signed resample query sequence components. PSD contrasts appear in Table 24.

Method / contrast	Remote (primary)	Neighborhood
Seven-layer cosine	0.5763	0.7513
Final-layer cosine	0.3696	0.6556
Fixed-PSD	0.6382	0.7705
ReaPO-PSD	**0.6446**	**0.7771**
Fixed-Signed	0.6090	0.7494
ReaPO-Signed	0.6399	0.7693
Signed − Fixed-Signed	$+0.0309$ $[0.0246, 0.0378]$	$+0.0199$ $[0.0124, 0.0293]$

CATH–BioLiP2 tests retrieval by shared structural-superfamily or ligand labels. Candidate selection uses neither model scores nor relevance labels and requires no opposing rankings. Neighborhood includes all eligible evaluation proteins except the query; Remote excludes its sequence component and retains one candidate per remaining component. Remote has 1,676 structural-superfamily queries and 1,806 ligand queries (Table 22). Each task uses checkpoints trained without its target relation family; the sets differ, and none jointly withholds both families (Appendix D.1).

On Remote, ReaPO-Signed reaches 0.6399 nDCG@10, compared with 0.5763 for seven-layer cosine. It improves over matched Fixed-Signed by **0.0309** (paired bootstrap 95% CI $[0.0246, 0.0378]$); the Neighborhood increment is **0.0199** ($[0.0124, 0.0293]$). Conditional PSD remains numerically higher on Remote at 0.6446; both Signed and PSD benefit from conditioning.

Signed gains over Fixed-Signed on Remote are $+0.0514$ for structural superfamily and $+0.0105$ for ligand identity. The overall gain is the mean of the two unrounded values. Table 27 reports these gains and positive gains for all three seeds. Per-relation results are descriptive, without additional significance tests.

External evaluation uses frozen checkpoints and fixed candidate and support sets, without retraining, checkpoint selection or support adjustment. Paired bootstrap intervals resample query sequence components, holding features, candidates and supports fixed.

5 Discussion and Conclusion

ReaPO-Signed learns relation-conditioned comparison rules over frozen protein features. The representation and support ablations show that both pretrained features and directed evidence contribute to its switching performance. One condition defines a common comparison rule across protein pairs; query and candidate identities do not enter condition induction.

Scope and limitations. Semantic families are operational annotation groups; their exclusion tests readout-family transfer, not unseen mechanisms or absence from PLM pretraining. The instruction-path ablation retains language-related supervision and applies to switching-trained readouts, not the target-only models. No global-sign external comparison has been run.

Crossed ranking and ordinary retrieval answer complementary questions: whether the scorer changes its ordering with the relation, and whether conditioning helps when candidates need not satisfy opposing preferences. The evidence supports relation-conditioned comparison within the tested protocols, rather than a universal advantage of signed geometry.

Reproducibility Statement

Implementation details, complete training objectives, data and split definitions, and evaluation protocols are provided in the main text and appendices.

AI Assistance Statement

The authors developed and verified all scientific content. Generative AI tools assisted with language editing, manuscript presentation, coding guidance, and checks of code and experiments.

References

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000. doi: 10.1038/75556.

Kosio Beshkov and Anders Malthe-Sorenssen. Towards understanding the shape of representations in protein language models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/384d189e40cf0486d3a4e304ffa12b6a-Abstract-Conference.html.

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.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021. doi: 10.1145/3458754.

David Ha, Andrew M. Dai, and Quoc V. Le. HyperNetworks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/pdf?id=rkpACe1lx.

Tymor Hamamsy, Meet Barot, James T. Morton, Martin Steinegger, Richard Bonneau, and Kyunghyun Cho. Learning sequence, structure, and function representations of proteins with language models. *bioRxiv*, pp. 2023.11.26.568742, 2023. doi: 10.1101/2023.11.26.568742.

Aashiq H. Kachroo, Jon M. Laurent, Christopher M. Yellman, Austin G. Meyer, Claus O. Wilke, and Edward M. Marcotte. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science*, 348(6237):921–925, 2015. doi: 10.1126/science.aaa0769.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

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

Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik L. L. Sonnhammer, Silvio C. E. Tosatto, Lisanna Paladin, Shriya Raj, Lorna J. Richardson, et al. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, 2021. doi: 10.1093/nar/gkaa913.

Dino Oglic and Thomas Gaertner. Learning in reproducing kernel krein spaces. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3859–3867, 2018. URL https://proceedings.mlr.press/v80/oglic18a.html.

Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz L. Pinto, Gustavo A. Salazar, Maxwell L. Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, et al. InterPro in 2022. *Nucleic Acids Research*, 51(D1):D418–D427, 2023. doi: 10.1093/nar/gkac993.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/37f65c068b7723cd7809ee2d31d7861c-Abstract.html.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118.

Robert Shaw, Samuel D. Love, and Claire D. McWhite. Evaluating pretrained protein language model embeddings as proxies for functional similarity. *Journal of Molecular Evolution*, 93(6): 765–776, 2025. doi: 10.1007/s00239-025-10282-4.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35:1026–1028, 2017. doi: 10.1038/nbt.3988.

Jin Su, Yan He, Shiyang You, Shiyu Jiang, Xibin Zhou, Xuting Zhang, Yuxuan Wang, Xining Su, Igor Tolstoy, Xing Chang, Hongyuan Lu, and Fajie Yuan. A trimodal protein language model enables advanced protein searches. *Nature Biotechnology*, 44(8):1310–1316, 2026. doi: 10.1038/s41587-025-02836-0.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023. doi: 10.1093/nar/gkac1052.

Sagar Vaze, Nicolas Carion, and Ishan Misra. GeneCIS: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6862–6872, 2023. doi: 10.1109/CVPR52729.2023.00663. URL https://openaccess.thecvf.com/content/CVPR2023/html/Vaze_GeneCIS_A_Benchmark_for_General_Conditional_Image_Similarity_CVPR_2023_paper.html.

Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 830–838, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Veit_Conditional_Similarity_Networks_CVPR_2017_paper.html.

Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. BERTology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.15222.

Vaishali P. Waman, Nicola Bordin, Andy Lau, Shaun Kandathil, Jude Wells, David Miller, Sameer Velankar, David T. Jones, Ian Sillitoe, and Christine Orengo. CATH v4.4: major expansion of CATH by experimental and predicted structural data. *Nucleic Acids Research*, 53(D1):D348–D355, 2025. doi: 10.1093/nar/gkae1087. URL https://doi.org/10.1093/nar/gkae1087.

Zeyuan Wang, Qiang Zhang, Keyan Ding, Ming Qin, Xiang Zhuang, Xiaotong Li, and Huajun Chen. InstructProtein: Aligning human and protein language via knowledge instruction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1114–1136. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.62. URL https://aclanthology.org/2024.acl-long.62/.

Ruben Weitzman, Peter Mørch Groth, Lood Van Niekerk, Aoi Otani, Yarin Gal, Debora Susan Marks, and Pascal Notin. Protriever: End-to-end differentiable protein homology search for fitness prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66397–66418, 2025. URL https://proceedings.mlr.press/v267/weitzman25a.html.

Minghao Xu, Xinyu Yuan, Santiago Miret, and Jian Tang. ProtST: Multi-modality learning of protein sequences and biomedical texts. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38749–38767, 2023. URL https://proceedings.mlr.press/v202/xu23t.html.

Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. Few-shot learning via embedding adaptation with set-to-set functions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8808–8817, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Ye_Few-Shot_Learning_via_Embedding_Adaptation_With_Set-to-Set_Functions_CVPR_2020_paper.html.

Chengxin Zhang, Xi Zhang, Lydia Freddolino, and Yang Zhang. BioLiP2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Research*, 52(D1): D404–D412, 2024. doi: 10.1093/nar/gkad630. URL https://doi.org/10.1093/nar/gkad630.

A Implementation and mathematical details

A.1 Architecture and initialization

The readout uses four support triples, a 64-dimensional condition and the seven ESM2 layers specified in Section 3. For a batch of b conditions, the anchored support contrasts have shape $b \times K \times 7 \times 1280$ before layer concatenation in Equation 1. Mean pooling excludes special tokens and each layer is normalized before concatenation. The frozen instruction representation has width 768. The support encoder has four attention heads of width 16, no positional encoding and zero attention dropout; its feed-forward block is $64 \to 128 \to 64$ with a Gaussian error linear unit (GELU). Each residual addition precedes layer normalization (LN). Table 5 collects the remaining shapes.

The support token is $t_k = \mathrm{LN}(W_e e_k)$, where $W_e \in \mathbb{R}^{64\times 8960}$ is bias-free and has row-normalized weights. Mean pooling the encoded tokens yields v. The instruction adapter applies layer normalization, an affine projection and GELU to yield u. The fusion map in Equation 2 is

$$f_\theta = \tanh \circ \mathrm{Linear}_{128\to 64} \circ \mathrm{GELU} \circ \mathrm{LN} \circ \mathrm{Linear}_{192\to 128}\,.$$

Table 5: **ESM2 readout architecture.** Affine maps include bias unless stated otherwise. Counts exclude frozen encoders and describe the complete readout. The fixed variants bypass the instruction and support paths (Appendix A.4).

Component	Specification	Parameters
Support basis and normalization	64×8960; LayerNorm(64)	573,568
Set encoder	Attention, feed-forward, two residual norms	33,472
Instruction adapter	LayerNorm(768), Linear(768,64), GELU	50,752
Condition fusion	$192 \to 128$, LN, GELU, $128 \to 64$, tanh	33,216
Diagonal basis	64×8960, no bias	573,440
Projection basis and head	64×8960; Linear(64,64)	577,600
Positive scales	Readout and training scale	2
Total trainable parameters	Complete readout	1,842,050

The support, diagonal and projection bases use normal initialization with standard deviation 0.02 and row normalization in the forward pass. Other affine and LayerNorm parameters use implementation defaults. Both coefficient heads scale logits by two. With logits $\ell_d \in \mathbb{R}^D$ and $\ell_\lambda \in \mathbb{R}^{64}$, Signed uses $d_{\mathcal{C}} = \rho_{\mathrm{S}}(\ell_d)$ and $\lambda_{\mathcal{C}} = \rho_{\mathrm{S}}(\ell_\lambda)$, where

$$\rho_{\mathrm{S}}(z) = \frac{\tanh z}{\max(\mathrm{mean}\,|\tanh z|, 10^{-6})}. \tag{9}$$

The diagonal head projects $c_{\mathcal{C}}$ through a learned row-normalized basis; the projection head is affine. PSD replaces ρ_{S} with $|\rho_{\mathrm{S}}|$. Mean-absolute normalization is applied separately to the two heads and can produce individual coefficients larger than one. The readout scale is positive and capped at 20. The squared feature norm in the score is floored at 10^{-8}.

Unrestricted scoring function. Unrestricted is not parameter-matched to the operator readout. The shared protein encoder is $p(x) = \mathrm{GELU}(\mathrm{LN}(W_p h(x) + b_p)) \in \mathbb{R}^{320}$. It also encodes each anchored contrast e_k before a $320 \to 256 \to 256$ support multilayer perceptron (MLP). A $768 \to 256 \to 256$ instruction MLP supplies another token. Learned instruction/support type embeddings precede one position-free, four-head Transformer block of width 256, feed-forward width 256 and zero dropout. With encoded tokens $z_I, z_1, \ldots, z_K$, the condition is $c = \mathrm{LN}(z_I + K^{-1}\sum_k z_k)$. Define $u(x) = \mathrm{GELU}(\mathrm{LN}(W_u p(x) + b_u)) \in \mathbb{R}^{256}$ and

$$t(q, x, c) = f([u(q); u(x); |u(q) - u(x)|; u(q) \odot u(x); c]),$$

where f is Linear$(1280, 128)$, LayerNorm, GELU and Linear$(128, 1)$. The pair logit is $\min(e^\gamma, 100)[t(q, x, c) - t(q, y, c)]$. This enforces candidate-swap antisymmetry; it does not require query–candidate symmetry or a bilinear factorization. For a fixed query and condition, candidates are ranked by t. All maps are trained under the matched crossed-ranking schedule.

Scoring factors. With projection width $R = 64$, define $A = [I_D; B]$ and $w_{\mathcal{C}} = [d_{\mathcal{C}}; \lambda_{\mathcal{C}}]$. The exact score factors are

$$\phi^q_{\mathcal{C}}(x) = \frac{\sqrt{\kappa}\ \mathrm{diag}(\sqrt{|w_{\mathcal{C}}|})Ah(x)}{n_{\mathcal{C}}(x)}, \qquad \phi^c_{\mathcal{C}}(x) = \mathrm{diag}(\mathrm{sign}\, w_{\mathcal{C}})\phi^q_{\mathcal{C}}(x). \tag{10}$$

Their dot product equals Equation 4, including its denominator floor; each has width $D + R = 9024$. PSD uses a common factor. For N candidates and Q queries, scoring is a $Q\times(D+R)$ by $(D+R)\times N$ matrix product. Candidate factors are cached within a condition and rebuilt when its instruction or support changes (Algorithm 1).

Measured implementation costs. For the switching-trained readouts on an idle A800 (FP32, TF32 disabled), a new condition with cached protein features for 10,000 candidates and 100 queries takes 92.60 ms for Signed and 654.40 ms for Unrestricted. The Unrestricted scorer reuses its condition-independent gallery cache. Candidate caches occupy 360.96 and 10.24 decimal MB, respectively. For one query, Unrestricted is faster: 8.05 versus 11.38 ms. Timings include condition/candidate preparation and cumulative query costs but exclude PLM extraction; cache sizes exclude shared features and model memory. These measurements characterize the stated workload and implementation.

A.2 Invariance, factorization and numerical limits

Permuting support rows commutes with row-wise maps and position-free self-attention. Mean pooling then removes the permutation, so the condition is support-order invariant in exact arithmetic. Query and candidate identities do not enter condition induction. Reversing support preferences negates each anchored contrast, but the nonlinear conditioner need not negate the final score.

Let $z(x) = \mathrm{diag}(\sqrt{|w_{\mathcal{C}}|})Ah(x)$ and $J_{\mathcal{C}} = \mathrm{diag}(\mathrm{sign}\, w_{\mathcal{C}})$. Then $z(q)^\top J_{\mathcal{C}} z(x) = h(q)^\top M_{\mathcal{C}} h(x)$ and $\|z(x)\|^2 = h(x)^\top H_{\mathcal{C}} h(x)$. Division by the explicitly floored $n_{\mathcal{C}}(x)$ proves Equation 10, including zero vectors. Because $\|z(x)\|/n_{\mathcal{C}}(x) \le 1$ and $\|J_{\mathcal{C}}\| \le 1$, $|s_{\mathcal{C}}(q,x)| \le \kappa$. Symmetry follows from $M_{\mathcal{C}} = M_{\mathcal{C}}^\top$; PSD additionally gives a positive-semidefinite Gram matrix for a fixed condition. Both variants define similarity scores; a distance requires a separate definition.

Generally $H_{\mathcal{C}} \neq |M_{\mathcal{C}}|$, where $|M_{\mathcal{C}}|$ denotes the spectral absolute value: projection rows need not be orthogonal. Instead, $|w_i| \pm w_i \ge 0$ implies $-H_{\mathcal{C}} \preceq M_{\mathcal{C}} \preceq H_{\mathcal{C}}$. Coefficient signs therefore describe the parameterization, not eigenvalues or identified biological factors. PSD itself can reverse rankings: with $q = (1,1)$, $a = (1,0)$ and $b = (0,1)$, normalized $\mathrm{diag}(2,1)$ scoring prefers a, whereas $\mathrm{diag}(1,2)$ prefers b. Our Signed–PSD result is an empirical constraint comparison, not an impossibility theorem for PSD.

Numerical convention. Equation 4 and the exact factors in Equation 10 floor squared weighted norms at 10^{-8}. The direct scoring path is used for crossed-ranking margins and the target-only Ordinary20 operator evaluations. The CATH–BioLiP2 retrieval implementation instead normalizes weighted factors with `F.normalize`, whose norm floor is 10^{-12}. These conventions agree in exact arithmetic when each nonzero weighted norm is at least 10^{-4}, and at zero, but can differ for smaller nonzero norms. The frozen evaluations retain their original arithmetic.

A frozen-checkpoint audit covers all four external operator variants: 72 checkpoints, 216 checkpoint–support contexts and all 2,354 proteins in the original feature bank (508,464 protein–context occurrences). No weighted norm is below 10^{-4}; the minimum is 0.5257 (also checked via $h(x)^\top H_{\mathcal{C}} h(x)$). Changing only the factor norm floor from 10^{-12} to 10^{-4} leaves the factors identical and changes no score or ranking on the original Remote and Neighborhood galleries. Thus the floor choice is inactive on these external evaluation inputs; no retraining or result replacement is involved.

This audit isolates the floor choice, not floating-point reduction order. Existing Ordinary20 direct-versus-coefficient-reuse checks have maximum absolute errors 4.77×10^{-7} for Signed and 2.09×10^{-7} for Fixed-Signed; external Signed/Fixed-Signed synthetic direct-versus-factorized checks have maximum error 1.43×10^{-6}. These separate checks do not establish bitwise equality between scoring implementations or extend the norm audit to internal evaluations and ablations.

A.3 Training objectives and shared settings

Only the readout is optimized. Let $g_{\mathcal{C}}(q,p,n) = \tau[s_{\mathcal{C}}(q,p) - s_{\mathcal{C}}(q,n)]$, where $\tau = \min(e^{\gamma}, 100)$ is a learned positive training scale. Write $\ell_{\delta}(z) = \mathrm{softplus}(\delta - z)$. A correct target margin is encouraged by $\ell_{.2}(g_{\mathcal{C}})$. For a corrupted condition $\widetilde{\mathcal{C}}$, preference for correct evidence is trained using $\ell_{\delta}(g_{\mathcal{C}} - g_{\widetilde{\mathcal{C}}})$. Corruptions are specified in the training data, not mined from evaluation outcomes.

Crossed-ranking training has three fixed stages. **Annotation training** uses broad relation-labelled triples, with

$$\mathcal{L}_A = \mathbb{E}_{\mathrm{fam}}\Big[\ell_{.2}(g_{\mathcal{C}}) + \tfrac{1}{2}\ell_{.2}(g_{\mathcal{C}} - g_{\widetilde{I},S}) + \tfrac{1}{2}\ell_{.2}(g_{\mathcal{C}} - g_{I,\widetilde{S}}) + \tfrac{1}{4}(1 - \cos(v,u))\Big] + \tfrac{1}{4}\mathcal{L}_{\mathrm{con}}. \quad (11)$$

$\mathbb{E}_{\mathrm{fam}}$ balances relation contributions within families. Wrong conditions come from a different relation chosen by instruction proximity in the training batch. $\mathcal{L}_{\mathrm{con}}$ is a supervised condition contrast: normalized fused conditions use temperature 0.1, excluding self-pairs; the numerator sums exponentiated similarities to same-relation conditions and the denominator sums over all other conditions. Anchors without a same-relation partner are excluded from this term.

Broad switch training uses natural contrasts with fixed q, x, y and two opposing criteria. Write $g_x = g_{\mathcal{C}_x}(q,x,y)$ and $g_y = g_{\mathcal{C}_y}(q,x,y)$, so the desired signs are positive and negative respectively. Let $z = (g_x, -g_y)$, and let $\widetilde{z}_j$ denote the corresponding corrupted-condition margins. The objective is

$$\mathcal{L}_X = \mathbb{E}_{\mathrm{pair}}\left[\tfrac{1}{2}\sum_{z_i \in z}\ell_{.2}(z_i) + \ell_{.2}(g_x - g_y) + \tfrac{1}{2}\sum_{j \in \{I,S,\mathrm{label}\}}\tfrac{1}{2}\sum_{i=1}^{2}\ell_{.1}(z_i - \widetilde{z}_{j,i})\right]. \quad (12)$$

The inner half averages the two directions; the outer half is each corruption's loss weight. **Switch refinement** retains Equation 12 and adds a weight-one, same-form preference loss for available annotation-mismatched support, averaged only over records with a valid replacement. Each update uses all eligible refinement-training records, with equal weight for each relation pair.

Annotation-mismatched support. In the refinement panel, correct support donors share a selected annotation with the target query–positive pair under the requested relation. The auxiliary preference assumes that demonstrations matching this annotation provide more relevant context within that relation. The control keeps the instruction, relation type, query, candidates and support size fixed, replacing the support with donors whose selected label is absent from every annotation of the target query; GO labels additionally obey the semantic-similarity cutoff defined in Appendix B.1. Donors obey the same protein-role disjointness rules. These replacements remain valid examples of the same relation: the loss encodes a training assumption about target relevance, not incorrect donor labels or identified causal mechanisms. This differs from wrong-support corruption, which uses the opposite relation's support, and label corruption, which reverses support preferences by swapping the positive and negative roles in each triple. Missing valid controls are masked in refinement and do not remove targets from the primary task. Labels guide construction but are not conditioner inputs. This refinement supervision is distinct from the primary crossed evaluation, where donors' shared labels are disjoint from the target pair's shared labels under each condition (Appendix B).

Optimization, schedules and matching. Both operator comparisons use AdamW with weight decay 0.01, gradient norm clipping at 5 and a fresh optimizer per stage. Seeds are 20260806, 20260807 and 20260808. Each arm uses all 19 eligible folds, withholding both endpoint families from every training stage, for 57 checkpoints per method. Within each comparison, records, supports, fold masks, seeded batch order and scheduled updates agree; the final scheduled checkpoint is used without evaluation-label selection. The ordinary trainer resets the seed after optional text encoding so its four arms start from identical shared weights. The global-sign head preserves the backbone initialization stream.

Table 6: **Crossed-ranking training schedule and loss weights.** LR: learning rate; T: target; X: condition switch; I/S: wrong instruction/support; L: reversed support preferences; M: available annotation-mismatched support; A/C: alignment/condition contrast.

Stage	Epochs	LR	Batch	T	X	I	S	L	M	A	C
Triplets	12	.001	256	1	–	.5	.5	–	–	.25	.25
Broad switches	12	.00075	256	1	1	.5	.5	.5	–	–	–
Refinement	160	.0005	Full	1	1	.5	.5	.5	1	–	–

Crossed training follows all three rows of Table 6. Ordinary retrieval uses the Triplets row's schedule but minimizes only $\mathbb{E}_{\text{fam}}\ell_{.2}(g_{\mathcal{C}})$; all switch, corruption, alignment and condition-contrast terms are absent. It uses recorded relation instructions without the crossed curriculum. Crossed training alternates generic and specific instructions in each stage, beginning with "Use the labelled protein-pair examples to infer the intended similarity criterion." The training log scale starts at 2 and the readout log scale at zero; their caps and coefficient initialization are specified above. Features and supports remain fixed under the instruction curriculum.

For N_A annotation triples and N_X broad switches after family removal, ordinary retrieval takes $12\lceil N_A/256\rceil$ updates, and crossed ranking adds $12\lceil N_X/256\rceil + 160$. Annotation and broad-switch stages contribute 1,248–1,728 and 468–1,416 updates per fold. Each refinement epoch makes one full-panel update with equal relation-pair weighting and valid-record averaging for the annotation-mismatched support term. Table 6 gives its weights; Equations 11–12 define the losses and margins.

A.4 Matched controls

Ordinary retrieval crosses coefficient constraint (Signed/PSD) with conditioning (full/fixed). Crossed ranking compares Signed with PSD plus a global sign. The shared feature bank, normalization, $K = 4$ support interface, 64-dimensional condition and rank-64 projection are unchanged. Appendix A.3 specifies optimization and budgets. Controls are matched within each task; crossed ranking and ordinary retrieval use different objectives.

Fixed-Signed and Fixed-PSD. Each fixed arm replaces both encoded inputs by zeros before fusion throughout training and inference:

$$u = v = 0, \qquad c_0 = f_\theta(0_{192}), \qquad M_0 = \operatorname{diag}(d(c_0)) + B^\top \operatorname{diag}(\lambda(c_0))B.$$

It is trained from initialization, not obtained by masking a fitted conditional model. Fusion biases, hidden normalization and output map, diagonal basis, projection head and basis, and positive readout/training scales remain learnable. The first fusion weight matrix receives no data gradient from zero inputs; the bypassed support encoder and instruction adapter receive no task gradient. Conditional counterparts retain the same shapes and normalization, but use $c_{\mathcal{C}} = f_\theta([v; u; v \odot u])$, with $\tanh$ for Signed and $|\tanh|$ for PSD. Fixed arms therefore have fewer active paths. They neither encode language nor use it to select training examples, and target-only training does not penalize their condition invariance.

PSD with a learned global sign. For the normalized conditional PSD score $s^{\text{PSD}}_{\mathcal{C}}$, including its positive scale, the 64-dimensional fused condition produces

$$\alpha_{\mathcal{C}} = \tanh(w^\top c_{\mathcal{C}} + b), \qquad s^{\text{global}}_{\mathcal{C}}(q, x) = \alpha_{\mathcal{C}} s^{\text{PSD}}_{\mathcal{C}}(q, x).$$

The continuous scalar multiplies the score after PSD normalization and is shared across all queries, candidates and coefficient directions under the condition. Its sign preserves or reverses the ranking, its magnitude scales margins, and zero yields ties. The head and PSD geometry are jointly trained through all three stages. Initialization is $w = 0, b = 1$, preserving the backbone random-number stream; the head adds 65 parameters. The underlying coefficients remain nonnegative, unlike coefficient-specific Signed scoring.

Crossed evaluation uses the fold withholding both evaluated families. Ordinary retrieval averages metrics over checkpoints withholding the target family; Appendix D.1 gives external target-specific routing.

A.5 Retrieval procedure

Shared projections cost $O(RD)$ per protein and condition-specific factors cost $O(D + R)$. A new condition rebuilds candidate factors; repeated queries amortize this cost. Here $h_{\text{text}}(I)$ denotes the cached, mean-pooled and normalized PubMedBERT instruction embedding.

Algorithm 1 Condition induction and retrieval with a frozen PLM

Require: instruction I, support $\{(a_k, p_k, n_k)\}_{k=1}^{K}$, queries $\mathcal{Q}$, gallery $\mathcal{G}$, trained readout θ

Retrieve normalized multilayer features $h(x)$ and shared projections $Bh(x)$.
for $k = 1, \ldots, K$ **do**
    $e_k \leftarrow h(a_k) \odot [h(p_k) - h(n_k)]$; $t_k \leftarrow \mathrm{LN}(W_e e_k)$.
end for
$v \leftarrow \mathrm{Mean}(\mathrm{SetEncoder}(\{t_k\}))$; $u \leftarrow \mathrm{TextAdapter}(h_{\text{text}}(I))$.
$c_{\mathcal{C}} \leftarrow f_\theta([v; u; v \odot u])$; obtain $d_{\mathcal{C}}, \lambda_{\mathcal{C}}$ with the trained variant's activation.
For every $x \in \mathcal{G}$, cache $\phi^c_{\mathcal{C}}(x)$ from Equation 10.
for $q \in \mathcal{Q}$ **do**
    Compute $\phi^q_{\mathcal{C}}(q)$; return candidates sorted by $\phi^q_{\mathcal{C}}(q)^\top \phi^c_{\mathcal{C}}(x)$.
end for

B Data and statistical protocol

B.1 Relations, annotations and eligibility

The registry uses frozen reviewed UniProt annotations and cross-references (The UniProt Consortium, 2023), including Gene Ontology (Ashburner et al., 2000), Pfam (Mistry et al., 2021) and InterPro (Paysan-Lafosse et al., 2023). Table 7 lists every canonical instruction; semantic families are operational holdout groups, and several relations share annotation sources.

Snapshot and annotation pool. Core20 uses UniProtKB/Swiss-Prot release 2026_02 (10 June 2026), retrieved on 1 August 2026. The frozen query selected reviewed entries with annotation score 5 and sequence length 40–1,022 residues. It returned 55,223 entries; removing 126 entries without any initial relation annotation left 55,097. Cross-references supply Pfam, InterPro, Gene3D and SUPERFAMILY annotations; curated fields supply EC, Rhea, ligand, location and membrane annotations. GO uses the 19 May 2026 ontology with `is_a`/`part_of` ancestry. Experimental GO evidence is restricted to EXP, IDA, IPI, IMP, IGI, IEP and their high-throughput counterparts HTP, HDA, HMP, HGI and HEP. Bound-ligand and cofactor annotations require experimental evidence used in manual assertion (ECO:0000269). Curated location and topology annotations are not restricted to experimental evidence. Supplementary mapping files and the final registry are identified by content hashes in the construction supplement; no live annotation lookup occurs during evaluation.

Internal sequence components and protein split. Internal grouping uses MMseqs2 settings `--min-seq-id 0.3 -c 0.8 --cov-mode 0 --cluster-mode 1`. Mode 0 applies the 80% coverage requirement to both sequences. We union the saved representative–member links with exact-sequence-hash matches, then name each component by its lexicographically smallest accession. Reconstructing these unions reproduces every component ID for the 38,957 proteins in the public annotation pool. These are the internal components, separately defined from the external evaluation graph.

The original component assignment uses seed 20260801 and 70/15/15 proportions; only the resulting 38,957-protein training partition enters this study's public annotation pool. This source partition is further split for the present study: deterministic component hashing with seed 20260806 assigns 20% to a held-out test split and the remainder to readout training; the exact hash convention is in `protocol_supplement/construction_parameters.json`. This yields 31,051 readout-training proteins in 10,787 components and 7,906 held-out test proteins in 2,690 components. Labels are retained using at least five proteins and five components in each of these two partitions. The remaining original partitions do not enter these internal experiments.

Table 7: **All 20 relations and their canonical instructions.** Identifiers are for this appendix only and are never model inputs. BP: biological process; Cell: cellular context; Dom: domain architecture; EC: enzyme class; Evol: evolutionary context; Struct: global structure; Lig: ligand context; Mem: membrane topology; MF: molecular function; Rxn: reaction chemistry.

ID	Family	Relation / canonical instruction
R01	Lig	**Bound ligand**. Find proteins sharing an experimentally supported bound ligand chemical identity.
R02	Lig	**Cofactor**. Find proteins sharing an experimentally supported cofactor chemical identity.
R03	EC	**EC2**. Find enzymes sharing the same first two levels of Enzyme Commission classification.
R04	EC	**EC3**. Find enzymes sharing the same first three levels of Enzyme Commission classification.
R05	EC	**EC4**. Find enzymes that are similar with respect to their complete EC activity.
R06	Struct	**Gene3D**. Find proteins sharing a Gene3D structural-domain superfamily.
R07	BP	**GO biological process**. Find proteins that are similar with respect to the biological processes they experimentally participate in.
R08	Cell	**GO cellular component**. Find proteins that are similar with respect to their experimentally supported cellular component or location.
R09	MF	**GO molecular function**. Find proteins that are similar with respect to their experimentally supported molecular function.
R10	Dom	**InterPro domain**. Find proteins sharing an integrated InterPro domain annotation.
R11	Evol	**InterPro superfamily**. Find proteins that are similar with respect to homologous superfamily membership.
R12	Evol	**Pfam clan**. Find proteins whose Pfam families belong to a common Pfam clan.
R13	Evol	**Pfam family**. Find proteins that are similar with respect to their Pfam family content.
R14	Rxn	**Rhea participant**. Find proteins whose curated Rhea reactions share an informative chemical participant.
R15	Rxn	**Rhea reaction**. Find proteins that are similar with respect to the biochemical reactions they catalyse.
R16	Cell	**Location: coarse**. Find proteins assigned to the same curated coarse subcellular compartment.
R17	Cell	**Location: exact**. Find proteins sharing the same exact reviewed UniProt subcellular-location term.
R18	Struct	**SUPERFAMILY**. Find proteins sharing a SUPERFAMILY structural classification.
R19	Evol	**Taxonomy family**. Find proteins from organisms that share the same NCBI Taxonomy family.
R20	Mem	**Membrane topology**. Find membrane proteins sharing the same broad transmembrane topology class.

From annotations to natural crossed candidates. For relations r_a, r_b, a query must have nonempty annotations for both. Candidate a shares an annotation with the query under r_a and is negative under r_b; candidate b satisfies the converse. A negative requires nonempty, disjoint label sets; for GO, Lin best-match-average similarity must additionally be at most 0.2. Lin similarity compares two GO terms using the information content of their most informative common ancestor. The symmetric best-match average takes each term's highest similarity to the other label set, averages within each direction, then averages the two directions. Missing or ambiguous labels are excluded.

The three proteins occupy distinct internal components. Candidate pairs match their query-relative taxonomic-distance bin, and their sequence lengths satisfy $\big|\,|\log(L_q/L_a)| - |\log(L_q/L_b)|\,\big| \le 0.5$. Taxonomy bins distinguish same species, close/mid/distant lineage within a superkingdom (shared-prefix fractions ≥ 0.75, $[0.5, 0.75)$ and < 0.5), different superkingdoms and unknown lineage; the matching constraint is disabled when taxonomy itself is the target criterion. The shared-prefix fraction is the number of matching initial lineage nodes divided by the shorter lineage length. Sampling uses deterministic relation-specific seeds and caps component reuse at three per relation pair in the natural-candidate source.

Relation-pair and fold eligibility. The construction screen contains 66 predefined candidate pairs: 42 core annotation pairs, 16 additional-annotation pairs and eight taxonomy pairs. A pair must yield at least seven crossed records in both the readout-training and held-out test partitions with four support triples per condition and disjoint role components. Each support anchor–positive pair must share an annotation under its condition. Annotation-mismatched support controls are optional and do

not determine primary eligibility. Forty-four pairs pass and 22 fall below this capacity criterion; eight failed pairs involve an axis outside the final 20-relation registry. Among the $\binom{20}{2} = 190$ possible Core20 pairs, 132 are outside the predefined screen, 14 fail its capacity rule and 44 are retained. The complete list distinguishes an untested combination from one that failed the screen. Pair eligibility uses annotation and support capacity, not model scores. The internal evaluations use the held-out test split, which was not used for readout training, method development, hyperparameter tuning or checkpoint selection.

For a target candidate in a retained pair, each condition receives four support triples drawn from natural crossed candidates of that same pair. The support therefore demonstrates the requested relation in the context of its competing relation. Each donor's shared labels under that condition must be disjoint from the target's shared labels. All 24 support proteins, across both conditions, and the query and two candidates occupy 27 distinct components. Support donors may share a non-target label. SHA256 ordering uses seed 20260806; donor reuse is capped at ten (observed maximum nine), and each pair contributes at most 100 targets and at least seven. Of 23,757 source candidates, 1,182 are outside the retained pair list. Among the remaining 22,575, the per-pair cap removes 18,193; none of the attempted uncapped targets fails donor construction. This gives 4,382 supported targets over 44 pairs. R08–R16 has 100 targets but both relations belong to cellular context, so it is excluded from strict two-family holdout. The remaining 4,282 targets constitute the 43 crossed pairs in Table 16.

Pairs with the same unordered endpoint-family pair share a training fold. A fold is eligible exactly when it contains a retained cross-family pair and nonempty training data after removing both endpoint families from every stage. All 19 supported folds satisfy this rule; the other 26 family combinations contain no retained pair (Table 9). No additional fold subset is selected. Across these folds, annotation, broad-switch and refinement pools contain 26,624–36,864, 9,779–30,058 and 668–2,557 records, respectively. Complete per-fold counts and overlap denominators are provided in `protocol_supplement/eligible_fold_counts.csv`.

Data for each training stage. Annotation training contains 2,048 episodes per relation, or 40,960 in total. It samples a retained label, a query–positive pair sharing that label, and a negative with disjoint annotations. Four support triples use other sampled labels where available; all 15 proteins occupy distinct components. Broad switch training attaches annotation-stage supports for the two relation criteria to 41,950 natural crossed targets, with all 27 role components distinct. Switch refinement uses 3,881 training records with four annotation-matched support triples per condition and the recorded annotation-mismatched supports used by its auxiliary loss. These supports come entirely from the training partition. Every stage removes all relations in both held-out families; annotation removal applies to its single relation, and switch removal applies to either relation endpoint. Target-ranking-only retrieval training uses the same annotation-stage pool and family exclusions for 12 epochs. Table 8 reports the distinct proteins actually present in these fixed records, rather than assuming every annotation-pool protein is used.

Three separate isolation checks. First, support and query/candidate components overlap in zero of the 4,282 crossed records. Second, query and candidate components overlap in zero records, and the two candidates also always occupy different components. Third, the union of all three training-stage records contains 30,706 proteins in 10,680 components, while crossed evaluation contains 5,712 proteins in 2,416 components. Their accession, identical-sequence and component intersections are all zero, both before family removal and in every actual fold. Ordinary20 contains 5,888 distinct proteins in 2,671 components including its support banks, and has zero overlap with annotation training at all three levels. Its per-relation support reservation and query–candidate checks are described below. Role isolation is local to a record or relation context: proteins can recur across records and relations.

The crossed result tests transfer to held-out relation families and protein components within the demonstrated contrast. It does not establish invariance to the competing relation, novel biological-label transfer or absence from frozen-encoder pretraining. Component separation is defined by the saved MMseqs2 graph and exact-sequence unions; it is not an exhaustive all-pairs guarantee against every remote homolog or shorter local alignment. Machine-readable lists of all 190 relation pairs and all 45 family combinations, per-fold overlap denominators, construction parameters and source hashes accompany the manuscript in `protocol_supplement/`.

Table 8: **Internal data and train–evaluation overlap.** Counts include query, candidate and support proteins; refinement also includes supplied annotation-mismatched supports. Training rows are counted before family removal. The last column gives protein/component overlap with the crossed evaluation. Each training stage also has zero identical-sequence overlap. Ordinary20 has zero overlap with annotation training at all three levels.

Data	Records	Proteins	Components	Overlap
Annotation training	40,960	30,676	10,668	0 / 0
Broad switch training	41,950	30,456	10,617	0 / 0
Switch refinement	3,881	17,208	7,862	0 / 0
Training union	–	30,706	10,680	0 / 0
Crossed evaluation	4,282	5,712	2,416	–
Ordinary20	1,280	5,888	2,671	–

Table 9: **All 45 family-pair combinations.** Entries count retained cross-family relation pairs; the symmetric matrix contains 19 nonzero unordered entries and 26 zeros. A zero means no retained relation pair supports that fold, not that all conceivable tasks for those families are impossible. Abbreviations follow Table 7.

Family	BP	Cell	Dom	EC	Evol	Struct	Lig	Mem	MF	Rxn
BP	–	3	1	0	2	2	0	0	1	0
Cell	3	–	1	5	5	4	2	1	3	3
Dom	1	1	–	0	0	0	0	0	1	0
EC	0	5	0	–	0	0	0	0	0	0
Evol	2	5	0	0	–	0	0	0	3	0
Struct	2	4	0	0	0	–	0	0	2	1
Lig	0	2	0	0	0	0	–	0	2	0
Mem	0	1	0	0	0	0	0	–	1	0
MF	1	3	1	0	3	2	2	1	–	0
Rxn	0	3	0	0	0	1	0	0	0	–

Ordinary20 uses a 7,906-entry held-out test manifest, reads only the requested relation's annotation field and omits proteins lacking it. Shared nonempty labels establish relevance. For GO, a nonoverlapping pair whose symmetric best-match Lin similarity exceeds 0.2 is ambiguous and excluded; otherwise it is negative. The GO ancestor sets and information-content values are precomputed and held fixed. These rules reduce annotation ambiguity but cannot establish exhaustive biological negatives.

Sampling uses deterministic SHA-256 orderings keyed by relation, role and namespace, with one representative per sequence component. Each seed reserves twelve distinct components for four support triples; the union across three seeds is excluded from query and candidate roles. Query and candidate proteins belong to different sequence components. Each gallery contains at least one positive and one negative, followed by deterministically selected and ordered unambiguous candidates. Every relation supplies 64 queries, each with 32 candidates, for 1,280 queries in total. Eligibility uses neither an opposite-answer requirement nor model outcomes. The positive/negative capacity rule differs from unconstrained database prevalence, and proteins may recur across relations or galleries.

All supported folds and three seeds contribute to the crossed evaluation, including the retrained input and operator ablations in Table 17.

B.2 METRICS AND UNCERTAINTY

For binary relevance y_j at rank j, $P = \sum_j y_j > 0$ and $k = \min(10, N)$,

$$\mathrm{nDCG@10} = \frac{\sum_{j=1}^{k} y_j / \log_2(j+1)}{\sum_{j=1}^{\min(k,P)} 1/\log_2(j+1)}, \qquad \mathrm{Hit@1} = y_1, \tag{13}$$

$$\mathrm{RR} = 1/\min\{j : y_j = 1\}, \qquad \mathrm{AP} = \frac{1}{P}\sum_{j=1}^{N} y_j \frac{\sum_{i=1}^{j} y_i}{j}. \tag{14}$$

Ties follow the original deterministic candidate index. Metrics, rather than scores, are averaged over eligible checkpoints/seeds for each query, then over queries within a relation, then equally over the 20 relations. Family plots average their constituent relations for display and need not reproduce the relation-weighted mean.

Crossed Joint requires $m_x > 0$ and $m_y < 0$ under Equation 8; equivalently, both target-oriented margins $(m_x, -m_y)$ must be positive. Ties fail. Directional accuracy counts the two directions separately, whereas Order tests $m_x > m_y$. A fixed score cannot satisfy both opposite preferences for unchanged proteins. The independent-choice value 0.25 is a descriptive reference, not a fitted null for correlated outputs.

Instruction and support preferences. Write $\mathcal{C}_x = (I_x, \mathcal{S}_x)$, $\mathcal{C}_y = (I_y, \mathcal{S}_y)$ and $\Delta(I, S) = s_{(I,S)}(q, x) - s_{(I,S)}(q, y)$. Set $u = \Delta(I_y, \mathcal{S}_x)$, $v = \Delta(I_x, \mathcal{S}_y)$ and use m_x, m_y from Equation 8. Record-level preferences are

$$p_I = \tfrac{1}{2}\big(\mathbf{1}[m_x > u] + \mathbf{1}[m_y < v]\big), \qquad p_S = \tfrac{1}{2}\big(\mathbf{1}[m_x > v] + \mathbf{1}[m_y < u]\big).$$

These swap only instructions or supports, respectively, holding proteins and checkpoint fixed. All records contribute; strict comparisons use no tolerance and give ties zero. For P_I, P_S, average records within each pair and seed, then the three seeds, then the 43 relation pairs equally; deterministic baselines contribute once.

Crossed contrasts average the three seeds within each relation pair and then weight all 43 pairs equally. For a pair-level method difference d_p, define $r_p = d_p - \bar{d}$. Let S_g sum residuals incident to node g, and T_h sum residuals in unordered node-pair intersection h. The symmetric node-cluster estimator is

$$\widehat{\mathrm{Var}}(\bar{d}) = \frac{1}{N^2}\left[\frac{G}{G-1}\sum_g S_g^2 - \frac{H}{H-1}\sum_h T_h^2\right], \qquad N = 43. \tag{15}$$

Relation-node cluster-robust variance estimation (CRVE) is primary ($G = 20, H = 43, t_{19}$); family-node sensitivity uses $G = 10, H = 19, t_9$. Family intersections sum their constituent relation-pair residuals before squaring, rather than treating the 19 overlapping folds as independent. Ablation inputs use recorded four-decimal seed means without intermediate rounding of method means.

Table 10: **Interval conventions.** Relation-node and family-node CRVE use the graph estimator above; bootstrap intervals retain paired method values. Corrections apply within the listed groups. The two Ordinary20 conditioning contrasts share one Bonferroni correction.

Comparison	Clustering / resampling	Coverage and correction
Signed − PSD + global sign	Relation nodes; family nodes (sensitivity)	95%; one contrast
Signed − support-selected seven-layer cosine	Relation nodes	98.3333%; Bonferroni, three
Signed − Unrestricted	Relation nodes	Nominal 95%
Pretrained − random features	Relation nodes	Nominal 95%
Complete − six retrained ablations	Relation nodes; family nodes (sensitivity)	99.1667%; Bonferroni, six
Six support-budget contrasts	Relation nodes	99.1667%; Bonferroni, six
Three frozen-sign contrasts	Relation nodes	98.3333%; Bonferroni, three
Individual support-budget methods	Relation nodes	Descriptive 95%
Ordinary20: Signed − Fixed-Signed	Paired semantic-family bootstrap	97.5%; Bonferroni, two
Ordinary20: PSD − Fixed-PSD	Paired semantic-family bootstrap	97.5%; Bonferroni, two
External: Signed − Fixed-Signed	Paired query-component bootstrap	Nominal 95% per track
External: PSD − Fixed-PSD	Paired query-component bootstrap	Nominal 95% per track

The three-comparison correction for the cosine diagnostic covers Signed minus (i) final-layer cosine with a support-selected sign, (ii) seven-layer cosine with a support-selected sign, and (iii) PSD with a learned global sign. The first row reports the global-sign contrast's nominal 95% interval separately; it is not the three-comparison adjusted interval. The frozen-sign family instead comprises sign removal, sign shuffling and random signs (Table 21).

Ordinary20 family bootstraps retain every constituent relation and recompute relation-macro differences on 20,000 paired draws. For joint reporting, we group the two conditioning contrasts in Table 14 and apply Bonferroni correction: the 1.25th and 98.75th percentiles give individual 97.5% intervals targeting simultaneous 95% coverage across the two contrasts. The Signed–Fixed-Signed and PSD–Fixed-PSD bootstrap seeds are 20260914 and 20260907, respectively. External query-component resampling is defined in Appendix D.1. Comparison-specific corrections do not establish a study-wide simultaneous error rate. Dir., Order, P_I/P_S, Hit@1, MRR and AP increments are descriptive unless explicitly stated otherwise.

Family-node analyses are post hoc sensitivity analyses on the same held-out test panel. Ten families, shared proteins and labels, and rounded inputs limit population inference, particularly near zero. Both clusterings are retained: the instruction-path contrast crosses zero under the primary relation-node interval despite a positive family-node interval. An interval crossing zero does not establish equivalence, and resampling pairs is not interchangeable with clustering shared nodes. The Signed–Unrestricted contrast, 0.071 $[-0.006, 0.148]$, meets the predeclared -0.02 noninferiority margin but does not establish superiority.

Table 11: **Strict crossed ranking.** Pair-macro means over 43 relation pairs and 19 held-out-family folds; trained scorers average three seeds. Learned global sign is a continuous condition-dependent scalar multiplying conditional PSD; Signed permits coefficient-specific signs. P_I/P_S: preference for correct instruction/support; strict ties give $P_I = 0$ for instruction-invariant controls.

Method	Joint	Dir.	Order	P_I	P_S
Support-selected global-sign diagnostics					
Final-layer cosine + support sign	0.305	0.530	0.305	0.000	0.305
Seven-layer cosine + support sign	0.458	0.616	0.458	0.000	0.458
Learned conditional scorers					
Unrestricted	0.477	0.680	0.730	0.649	0.714
ReaPO-PSD	0.130	0.543	0.727	0.690	0.713
ReaPO-PSD + global sign	0.514	0.695	0.734	0.693	0.718
ReaPO-Signed	**0.548**	**0.718**	**0.747**	**0.719**	**0.725**

B.3 FROZEN REPRESENTATION CONFIGURATIONS

The representation comparisons in Table 2 use separately trained Signed readouts with the same three-stage objectives, 19 strict-L2FO folds and three training seeds. Protein encoders remain frozen. Residue states are mean-pooled after excluding special tokens; each layer vector is ℓ_2-normalized before concatenation. ESM2 uses layers $(0, 6, 12, 18, 24, 30, 33)$, each of width 1,280, giving $D = 8{,}960$. Layer zero denotes the embedding input.

ProtT5 and ProtBERT. ProtT5 uses `Rostlab/prot_t5_xl_uniref50` encoder layers $(0, 4, 8, 12, 16, 20, 24)$. ProtBERT uses `Rostlab/prot_bert` (UniRef100) layers $(0, 5, 10, 15, 20, 25, 30)$. Both supply seven 1,024-dimensional vectors, giving $D = 7{,}168$ with the same per-layer normalization. Inputs are uppercase, space-separated residues with U/Z/O/B mapped to X. Cached representations are stored in float16 and loaded in float32 for normalization and readout training. The readout's input-dependent parameter shapes follow each encoder's width; the training protocol and operator form are shared.

Random-encoder control. The control instantiates the ESM2 architecture with 33 layers, hidden width 1,280, 20 attention heads, the ESM-1b alphabet and token dropout, without loading pretrained tensors. Representation seeds 120260806, 120260807 and 120260808 correspond respectively to readout-training seeds 20260806, 20260807 and 20260808. It uses the same selected layers, residue pooling and normalization as pretrained ESM2.

CountSketch control. The parameter-free input contains normalized contiguous 1-, 2- and 3-mer frequencies and $\log(1 + L)/\log(1 + L_{\max})$, where $L_{\max}$ is the maximum length in the frozen sequence manifest. The alphabet is `ACDEFGHIKLMNPQRSTVWYX`; other residues map to X. Each k-mer count is divided by $L - k + 1$. Seven deterministic signed sketches of width 1,280 give the same $D = 8{,}960$ interface as ESM2. Each sketch sums input features into hash-selected bins with fixed signs, then is normalized before concatenation. Hash inputs are the view index, corresponding ESM2 layer index and feature index; BLAKE2b uses a 16-byte digest and fixed personalization `EvoCrossPVC-R1`. The first eight digest bytes (little-endian, modulo 1,280) select the bin and the low bit of the ninth byte selects the sign. There is no sampled sketch seed, learned feature extractor or annotation input. The supplied `protocol_supplement/representation_config.json` records the exact indexing, hash payload, feature-artifact hashes and source bindings.

B.4 BASELINE INPUTS

Equal-seven cosine averages separately normalized layer cosines. Support ridge normalizes concatenated protein features and their pair products, fits eight $+1/-1$ pairs induced by four triples without an intercept, and selects $\lambda \in \{.01, .1, 1, 10, 100\}$ by leave-one-anchor-out support mean squared error (MSE). Ties within 10^{-12} prefer larger λ. Neither instructions nor query labels enter this fit.

Support-selected sign uses $\alpha_{\mathcal{C}} = \operatorname{sign}_+(\sum_k [s_0(a_k, p_k) - s_0(a_k, n_k)])$ and ranks by $\alpha_{\mathcal{C}} s_0(q, x)$, with zero mapped to $+1$. It uses support margins only, not instructions, relation-name lookup or query labels. The learned global-sign control rescales or reverses conditional PSD with a continuous signed scalar, without coefficient-specific signs. Within each checkpoint, Fixed-Signed and Fixed-PSD each share one learned operator across conditions; their conditional counterparts generate it from $I + S$. Appendix A.4 specifies the trainable parameters, scalar gate and comparison-specific training settings.

Unrestricted receives the same bank, language and support as the operators. ProtST official cosine and ProTrek direct sequence are condition-free sequence baselines. ProtST/ProTrek adapters use official frozen sequence/text representations and the Ordinary20 evaluation protocol defined here.

Instruction adapters. The ProtST and ProTrek adapters in Table 3 reuse three-stage instruction-only checkpoints. Their frozen encoders and shared feature-wise linear modulation (FiLM) heads are specified below; the four ordinary-retrieval operator arms use separate target-only training.

Configuration	ProtST adapter	ProTrek adapter
Frozen weights	`protst_esm2.pth`	`ProTrek_650M.pt`
SHA256 prefix	`44485e44c8ee`	`578b1e15d1a9`
Protein pooling	Residue mean, then `graph_mlp`	ESM2 first-token projection
Text pooling	Non-special-token mean, then `text_mlp`	PubMedBERT first-token projection
Input width / hidden width	512 / 2,834	1,024 / 1,867

Both input embeddings are L2-normalized. The protein head is $\mathrm{LN}(d) \to \mathrm{Linear}(d, w) \to \mathrm{GELU} \to \mathrm{Linear}(w, 512) \to \mathrm{LN}(512)$; the text head is $\mathrm{LN}(d) \to \mathrm{Linear}(d, 64) \to \mathrm{GELU}$. Two zero-initialized, bias-free $64 \to 512$ maps produce γ_I and β_I. For protein-head output $z(x)$, FiLM gives

$$\widehat{z}_I(x) = \mathrm{normalize}_{\ell_2}((1 + \gamma_I) \odot z(x) + \beta_I)\,.$$

Thus initialization preserves $z(x)$ before normalization. The same conditioned mapping is applied to the query and each candidate, and the resulting vectors are compared by cosine similarity; support examples are unused.

Training excludes both held-out families in all 19 folds, with seeds 20260806–20260808. Annotation and broad-switch stages each use 12 epochs and batches of 128; refinement uses 160 full-panel updates. AdamW uses learning rates 10^{-3}, 7.5×10^{-4} and 5×10^{-4}, weight decay 0.01 and gradient-norm clipping at 5, resetting optimizer state between stages. The final scheduled checkpoint is retained.

Annotation training combines target ranking, wrong-instruction preference (weight 0.5) and a sigmoid contrastive loss (weight 1). For each query, the contrastive term selects up to 32 highest-scoring eligible in-batch negatives, requiring at least eight. It excludes same-relation or same-label pairs, duplicate positives and instruction cosine ≥ 0.98. The positive-pair loss is added to the mean loss over selected negatives, and query losses are aggregated with family balancing. Switch stages use correct-side and opposing-margin terms (weight 1 each) plus wrong-instruction preference (0.5), with target/switch margin 0.2 and corruption margin 0.1. All support-related losses are zero. Complete settings and weight hashes are in `protocol_supplement/instruction_adapter_config.json`.

C Additional results and interpretation

C.1 Ordinary20 per-relation results

Table 12 expands the retrieval results in Table 3. The operator columns use the same recorded relation summaries as Figure 4, including Fixed-PSD. Their differences include both gains and declines.

Table 12: **Ordinary20 per-relation nDCG@10.** All 20 relations, in registry order. The four operator columns use target-only training. Values are rounded to three decimals.

Relation	Cosine-7	ProtST	Fixed-S	Signed	Fixed-P	PSD
R01 Bound ligand	0.357	0.407	0.410	0.464	0.436	0.461
R02 Cofactor	0.395	0.467	0.470	0.485	0.445	0.438
R03 EC2	0.264	0.381	0.378	0.352	0.280	0.274
R04 EC3	0.373	0.550	0.593	0.561	0.534	0.551
R05 EC4	0.546	0.779	0.837	0.849	0.789	0.807
R06 Gene3D	0.534	0.517	0.724	0.788	0.801	0.859
R07 GO biological process	0.320	0.398	0.315	0.316	0.297	0.291
R08 GO cellular component	0.486	0.540	0.473	0.520	0.492	0.507
R09 GO molecular function	0.429	0.524	0.475	0.469	0.458	0.476
R10 InterPro domain	0.559	0.764	0.818	0.849	0.865	0.861
R11 InterPro superfamily	0.639	0.639	0.797	0.816	0.874	0.897
R12 Pfam clan	0.540	0.548	0.679	0.724	0.754	0.781
R13 Pfam family	0.712	0.848	0.880	0.898	0.899	0.891
R14 Rhea participant	0.279	0.429	0.342	0.349	0.347	0.342
R15 Rhea reaction	0.614	0.711	0.757	0.777	0.751	0.775
R16 Location: coarse	0.511	0.576	0.518	0.552	0.523	0.548
R17 Location: exact	0.403	0.552	0.459	0.461	0.388	0.382
R18 SUPERFAMILY	0.540	0.543	0.709	0.773	0.811	0.864
R19 Taxonomy family	0.370	0.347	0.302	0.339	0.349	0.358
R20 Membrane topology	0.573	0.550	0.509	0.575	0.607	0.617
Relation macro	0.472	0.554	0.572	0.596	0.585	0.599

Fixed-S: Fixed-Signed; Fixed-P: Fixed-PSD. Family means average relation values; the macro row uses each method's relation-macro aggregate. Rounding can affect averages computed from displayed cells.

Table 13: **Complete Ordinary20 aggregate metrics.** The four operators use target-only training; scores are rounded to three decimals. Sequence baselines and adapters have the inputs and training described in Appendix B.

Method	Inputs	nDCG@10	Hit@1	MRR	AP
ESM2 final cosine	Seq.	0.367	0.278	0.423	0.343
ESM2 seven-layer cosine	Seq.	0.472	0.398	0.529	0.441
ProtST official cosine	Seq.	0.554	0.437	0.585	0.479
ProTrek direct sequence	Seq.	0.466	0.379	0.520	0.424
ESM2 support ridge	Seq.+S	0.229	0.145	0.283	0.232
ProtST adapter	Seq.+I	0.508	0.438	0.570	0.478
ProTrek adapter	Seq.+I	0.512	0.427	0.560	0.478
Fixed-Signed	Seq.	0.572	0.500	0.622	0.541
ReaPO-Signed	Seq.+I+S	0.596	0.516	0.637	0.562
Fixed-PSD	Seq.	0.585	0.520	0.637	0.554
ReaPO-PSD	Seq.+I+S	0.599	0.525	0.644	0.562

Table 14: **Ordinary20 paired conditioning contrasts in nDCG@10.** Bonferroni-adjusted simultaneous 95% intervals for the two contrasts. Each individual interval uses the 1.25th and 98.75th percentiles (97.5%) of 20,000 paired family-bootstrap draws over ten families, retaining relation-macro weighting. The PSD interval uses recorded three-decimal relation means.

Contrast	Δ	Adjusted interval
Signed − Fixed-Signed	$+0.0236$	$[0.0049, 0.0411]$
PSD − Fixed-PSD	$+0.014$	$[0.006, 0.027]$

C.2 Crossed-ranking coverage and sensitivity

For relation pair p, $\Delta_p = \bar{J}_{\text{Signed},p} - \bar{J}_{\text{PSD+GS},p}$ uses means over three training seeds. The primary mean weights all 43 pairs equally. Family-pair means average constituent pairs; their equal-weight macro gives the alternative summary over 19 family pairs. A family deletion removes all incident evaluation pairs (the union for Cell+MF), then averages remaining pairs equally. Models, supports and training holdouts stay fixed; all analyses here are descriptive.

Family pair	Signed	PSD+GS	Pairs	N	Gain (percentage points)
BP–Cell	0.1878	0.1767	3	300	+1.11
BP–Dom	0.9100	0.8933	1	100	+1.67
BP–Evol	0.7483	0.7334	2	200	+1.50
BP–MF	0.2667	0.2067	1	100	+6.00
BP–Struct	0.7533	0.7184	2	200	+3.50
Cell–Dom	0.8900	0.8567	1	100	+3.33
Cell–EC	0.5294	0.4953	5	486	+3.41
Cell–Evol	0.6953	0.6220	5	500	+7.33
Cell–Lig	0.2150	0.1967	2	200	+1.83
Cell–MF	0.2156	0.1856	3	300	+3.00
Cell–Mem	0.2767	0.2800	1	100	-0.33
Cell–Rxn	0.4811	0.4444	3	300	+3.67
Cell–Struct	0.7908	0.7242	4	400	+6.67
Dom–MF	0.8200	0.8000	1	100	+2.00
Evol–MF	0.7589	0.7456	3	300	+1.33
Lig–MF	0.2102	0.1904	2	196	+1.99
Mem–MF	0.1933	0.2000	1	100	-0.67
Struct–MF	0.7617	0.7500	2	200	+1.17
Struct–Rxn	0.6300	0.5467	1	100	+8.33

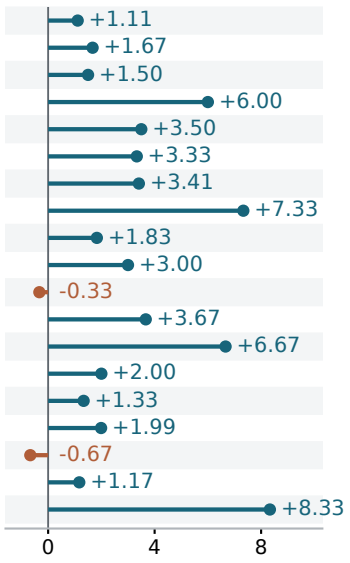


Figure 5: **Signed gains across all 19 family pairs.** Accuracy and gain columns average the four-decimal entries in Table 16 equally within each family pair. PSD+GS is conditional PSD with a learned global sign; N counts crossed records once. Gains are expressed in percentage points.

Table 15: **Weighting and family-deletion sensitivity.** Joint accuracy (both columns to four decimals) and gain use the four-decimal entries in Table 16. $P/F/N$: remaining relation pairs/family pairs/records; $+/-$: positive/negative pair differences. Deletions affect evaluation aggregation only.

Aggregation / excluded family	Signed	PSD+GS	Δ	P	F	N	$+/-$
Relation-pair equal (primary)	0.5484	0.5140	+0.0344	43	19	4,282	37/6
Family-pair equal	0.5439	0.5140	+0.0299	43	19	4,282	37/6
Remove incident relation pairs; equal weight among remaining pairs							
BP	0.5540	0.5167	+0.0374	34	14	3,382	30/4
Cell	0.6277	0.6042	+0.0235	16	10	1,596	12/4
Dom	0.5240	0.4888	+0.0353	40	16	3,982	34/6
EC	0.5509	0.5164	+0.0345	38	18	3,796	32/6
Evol	0.4949	0.4633	+0.0316	33	16	3,282	29/4
Struct	0.4929	0.4624	+0.0305	34	15	3,382	28/6
Lig	0.5828	0.5468	+0.0360	39	17	3,886	34/5
Mem	0.5637	0.5273	+0.0363	41	17	4,082	37/4
MF	0.5811	0.5407	+0.0405	30	12	2,986	27/3
Rxn	0.5515	0.5185	+0.0330	39	17	3,882	33/6
Cell + MF	0.7572	0.7239	+0.0333	6	4	600	5/1

Gains are positive in 37/43 relation pairs and 17/19 family pairs; the median pair gain is 0.0300 (range -0.0133 to $+0.1200$). The Cell+MF deletion retains only six pairs and 600 records (five positive, one negative), so its effect describes that narrower population. Family abbreviations follow Table 7.

Table 16: **All 43 relation-pair contrasts.** Strict Joint accuracy, averaged over three training seeds. Both accuracy columns are reported to four decimal places. PSD+GS is conditional PSD with a learned global sign. Δ is Signed minus PSD+GS. N is the crossed-record count, not multiplied by seeds. IDs refer to Table 7.

Relation pair	Family pair	Signed	PSD+GS	Δ	N
R01–R08	Cell–Lig	0.2133	0.1900	+0.0233	100
R01–R09	Lig–MF	0.2433	0.1967	+0.0467	100
R02–R08	Cell–Lig	0.2167	0.2033	+0.0133	100
R02–R09	Lig–MF	0.1771	0.1840	−0.0069	96
R03–R08	Cell–EC	0.2933	0.2367	+0.0567	100
R04–R08	Cell–EC	0.5133	0.4700	+0.0433	100
R04–R17	Cell–EC	0.5767	0.5533	+0.0233	100
R05–R08	Cell–EC	0.6705	0.6434	+0.0271	86
R05–R16	Cell–EC	0.5933	0.5733	+0.0200	100
R06–R07	BP–Struct	0.7433	0.7067	+0.0367	100
R06–R08	Cell–Struct	0.8000	0.7867	+0.0133	100
R06–R09	Struct–MF	0.7400	0.7333	+0.0067	100
R06–R14	Struct–Rxn	0.6300	0.5467	+0.0833	100
R06–R16	Cell–Struct	0.7667	0.6633	+0.1033	100
R07–R08	BP–Cell	0.1700	0.1567	+0.0133	100
R07–R09	BP–MF	0.2667	0.2067	+0.0600	100
R07–R10	BP–Dom	0.9100	0.8933	+0.0167	100
R07–R11	BP–Evol	0.7933	0.7567	+0.0367	100
R07–R12	BP–Evol	0.7033	0.7100	−0.0067	100
R07–R16	BP–Cell	0.2133	0.1900	+0.0233	100
R07–R17	BP–Cell	0.1800	0.1833	−0.0033	100
R07–R18	BP–Struct	0.7633	0.7300	+0.0333	100
R08–R09	Cell–MF	0.2200	0.2100	+0.0100	100
R08–R10	Cell–Dom	0.8900	0.8567	+0.0333	100
R08–R11	Cell–Evol	0.7700	0.7400	+0.0300	100
R08–R15	Cell–Rxn	0.5300	0.5033	+0.0267	100
R08–R18	Cell–Struct	0.8533	0.8200	+0.0333	100
R08–R20	Cell–Mem	0.2767	0.2800	−0.0033	100
R09–R10	Dom–MF	0.8200	0.8000	+0.0200	100
R09–R11	Evol–MF	0.7333	0.7200	+0.0133	100
R09–R12	Evol–MF	0.6767	0.6900	−0.0133	100
R09–R13	Evol–MF	0.8667	0.8267	+0.0400	100
R09–R16	Cell–MF	0.2500	0.2067	+0.0433	100
R09–R17	Cell–MF	0.1767	0.1400	+0.0367	100
R09–R18	Struct–MF	0.7833	0.7667	+0.0167	100
R09–R20	Mem–MF	0.1933	0.2000	−0.0067	100
R12–R16	Cell–Evol	0.7167	0.6433	+0.0733	100
R13–R16	Cell–Evol	0.9100	0.8100	+0.1000	100
R13–R17	Cell–Evol	0.8433	0.7233	+0.1200	100
R14–R16	Cell–Rxn	0.3267	0.2867	+0.0400	100
R15–R16	Cell–Rxn	0.5867	0.5433	+0.0433	100
R16–R18	Cell–Struct	0.7433	0.6267	+0.1167	100
R17–R19	Cell–Evol	0.2367	0.1933	+0.0433	100
All 43 pairs	Pair macro	0.5484	0.5140	+0.0344	4,282

These summaries average four-decimal records and round only for final display (half-up); unrounded inputs can give different last digits. Data: `coverage_supplement/`.

C.3 Ablations, support size and coefficient signs

These experiments use three-stage switching-trained readouts. The Table 17 ablations change the support representation or scoring form. Without anchor uses $h(p) - h(n)$, positive-only uses $h(a) \odot h(p)$, and negative-only uses $-h(a) \odot h(n)$. Diagonal-only sets the low-rank coefficients to zero; low-rank-only sets the diagonal coefficients to zero, during both training and evaluation. Retraining allows the remaining components to compensate. The contrasts therefore measure restricted-model performance, not a unique decomposition of biological mechanisms.

Instruction-path ablation. Without instruction path masks the instruction and instruction–support interaction inputs to fusion, while retaining language-related losses and negative-condition sampling. Instruction invariance gives $P_I = 0$ under strict ties; exchanging the paired support conditions gives $P_S = \mathrm{Order}$. This tests the scoring path under retraining, not the removal of all language supervision.

Table 17: **Retrained input and operator ablations.** All 43 relation pairs, 19 folds and three seeds. Table 18 gives adjusted Joint contrasts under relation-node and family-node clustering. Other metrics are descriptive.

Variant	Joint	Dir.	Order	P_I	P_S
ReaPO-Signed	**0.548**	**0.718**	**0.747**	**0.719**	**0.725**
Input ablations					
Without instruction path	0.539	0.690	0.724	0.000	0.724
Without anchor	0.326	0.591	0.635	0.700	0.525
Positive-only support	0.434	0.670	0.718	0.677	0.681
Negative-only support	0.385	0.646	0.693	0.671	0.625
Operator ablations					
Diagonal only	0.523	0.697	0.720	0.690	0.715
Low-rank only	0.536	0.707	0.746	0.704	0.724

Table 18: **Adjusted Joint contrasts for the retrained ablations in Table 17.** Complete minus variant over 43 pairs, 19 folds and three seeds, using aggregates before display rounding. Six-contrast Bonferroni 99.1667% intervals: primary relation-node CRVE (20 nodes, t_{19}) and family-node sensitivity (10 nodes, t_9).

Complete minus variant	ΔJoint	Relation-node CI	Family-node CI
Without instruction path	+0.0091	$[-0.0063, 0.0246]$	$[0.0005, 0.0177]$
Without anchor	+0.2221	$[0.1148, 0.3294]$	$[0.0959, 0.3483]$
Positive-only support	+0.1146	$[0.0680, 0.1613]$	$[0.0579, 0.1714]$
Negative-only support	+0.1635	$[0.0740, 0.2530]$	$[0.0805, 0.2465]$
Diagonal only	+0.0254	$[0.0022, 0.0486]$	$[0.0004, 0.0505]$
Low-rank only	+0.0126	$[0.0033, 0.0220]$	$[0.0015, 0.0238]$

The family-node lower bound for adding the low-rank term is close to zero; the table retains both clusterings.

Varying support size without retraining. Increasing inference support from one to four triples improves all three frozen scorers, with positive corrected intervals; these are not retrained few-shot regimes.

Table 19: **Inference-support budget.** Existing frozen L2FO models are evaluated with $K = 1, 2, 4$; these are not separately trained few-shot regimes. Intervals are descriptive 95% relation-node intervals.

Model	K	Joint	Dir.	Joint interval
ReaPO-Signed	1	0.458	0.663	[0.375, 0.541]
ReaPO-Signed	2	0.509	0.695	[0.410, 0.608]
ReaPO-Signed	4	0.548	0.718	[0.437, 0.660]
ReaPO-PSD	1	0.107	0.530	[0.097, 0.118]
ReaPO-PSD	2	0.115	0.534	[0.101, 0.129]
ReaPO-PSD	4	0.130	0.543	[0.114, 0.146]
Unrestricted	1	0.423	0.653	[0.324, 0.523]
Unrestricted	2	0.461	0.670	[0.354, 0.567]
Unrestricted	4	0.477	0.680	[0.370, 0.585]

Table 20: **All six support-budget contrasts.** Relation-node intervals use Bonferroni 99.1667% coverage across six comparisons. Differences use unrounded method means.

Comparison	ΔJoint	Corrected interval
Signed / K4-K1	0.090	[0.049, 0.132]
Signed / K4-K2	0.039	[0.013, 0.065]
PSD / K4-K1	0.022	[0.011, 0.033]
PSD / K4-K2	0.014	[0.007, 0.021]
Unrestricted / K4-K1	0.054	[0.027, 0.081]
Unrestricted / K4-K2	0.017	[0.005, 0.028]

Changing coefficient signs without retraining. These interventions preserve coefficient magnitudes and change only their signs in switching-trained Signed. Crossed Joint is 0.548 with original signs, 0.023 after sign removal, 0.431 after shuffling and 0.115 with random signs. Table 21 gives corrected contrasts. These interventions change coefficients at inference, rather than retraining a constrained model; their conclusions concern the switching-trained readout.

Table 21: **Frozen-checkpoint sign contrasts.** Crossed intervals have 98.3333% relation-node coverage (three contrasts). All contrasts report original minus intervention; estimates are rounded for display.

Comparison	Panel / direction	Δ	Corrected interval
remove sign	Crossed: original minus intervention	0.526	[0.384, 0.667]
shuffle sign	Crossed: original minus intervention	0.117	[0.064, 0.170]
random sign	Crossed: original minus intervention	0.434	[0.297, 0.570]

D External retrieval without conflict selection

D.1 Ordinary retrieval without conflict selection

The external panel uses CATH v4.4 structural-superfamily labels (Waman et al., 2025) and BioLiP2 ligand identities (Zhang et al., 2024). The source pool contains 3,048 proteins with a single label for each relation. Evaluation retains 2,286 proteins and 68 support proteins under the existing sequence-component split. Components are defined by MMseqs clustering at 30% sequence identity and 80% coverage. Components overlapping readout-training data are excluded, and support and evaluation components are disjoint. Observed ligand identity defines relevance, not binding affinity or mutually exclusive binding.

Tracks and sample sizes. Remote is the primary analysis; Neighborhood assesses sensitivity to the candidate pool. Both restrict query–candidate length ratios to at most 1.5. Neighborhood includes all eligible evaluation sequences other than the query itself. Remote excludes the query component. It orders the remaining candidates by a deterministic hash of the query and candidate identifiers, retaining the first candidate from each component. Gallery construction uses neither model scores nor relevance labels and imposes no conflict requirement. A query is evaluated for a relation only when its fixed gallery contains at least one positive and one negative. Eligibility is assessed separately for the two relations (Table 22).

Table 22: **External evaluation population.** Query counts are relation-specific; a protein may contribute to both relations. G counts distinct sequence components across all eligible query–relation units in each track; these components are resampled jointly.

Track	CATH queries	Ligand queries	Components (G)
Remote (primary)	1,676	1,806	996
Neighborhood (sensitivity)	2,029	2,172	1,017

Training-fold and checkpoint correspondence. Signed, Fixed-Signed, PSD and Fixed-PSD use the target-specific checkpoint routing in Table 23, with seeds 20260806, 20260807 and 20260808. Each method contributes 12 checkpoints to structural retrieval and six to ligand retrieval, for 18 distinct checkpoints in total. Every checkpoint excludes both families named by its training fold, but no fold in this evaluation jointly excludes the structure and ligand families. The two target relations therefore use different checkpoint sets: these results assess transfer to each held-out target family, not simultaneous transfer to both families from one jointly held-out model. All checkpoints use the target-only ordinary training specified in Appendix A.3; external data do not enter retraining or checkpoint selection. All four methods use the same three $K = 4$ support panels; fixed readouts ignore their contents. Within each query–relation unit, metrics are averaged over support panels, eligible folds and then seeds before the query and relation aggregation below.

Table 23: **Target-family holdout for external operator retrieval.** Signed, Fixed-Signed, PSD and Fixed-PSD use these folds. Each fold withholds the target family and one companion family. Counts are per method and include all three seeds.

Held-out target	Other held-out family	Folds	Checkpoints
Structure (CATH)	Biological process; cellular context; molecular function; reaction chemistry	4	12
Ligand (BioLiP2)	Cellular context; molecular function	2	6

Matched comparison and uncertainty. Fixed-Signed and Fixed-PSD learn condition-independent geometry; their conditional counterparts use instructions and directed support. After the within-query averaging described above, metrics are averaged over queries within each relation. The reported nDCG@10 gives the two relations equal weight.

For Signed-minus-Fixed-Signed, both tracks use a paired cluster bootstrap: sample G query sequence components with replacement, retaining all eligible relation units and the paired method values within each sampled component. Each replicate recomputes the equal-relation macro difference. We use 20,000 draws with base seed 20260914 and deterministic track-specific random streams. Linear-interpolated 2.5th and 97.5th percentiles yield the nominal 95% intervals in Table 4. CATH-superfamily and ligand-identity clusterings are separate sensitivity definitions; the reported intervals use sequence components. The feature bank, galleries and support are held fixed, so these intervals exclude support-bank and gallery-population uncertainty. They are not simultaneous intervals for all method comparisons.

Signed uses separate query and candidate factors as in Equation 10; a common Gram factor would erase coefficient signs. Fixed-Signed generates one learned operator per model, so its scores do not depend on the support panel.

Signed-minus-Fixed-Signed has positive paired intervals under all three cluster definitions on both tracks (Table 26). Its descriptive increments are also positive for each relation axis and each seed (Table 27). PSD-minus-Fixed-PSD also has positive paired intervals (Table 24). These within-parameterization effects answer the conditioning question; they do not imply that every conditional scorer outperforms every fixed geometry.

Table 24: **External PSD conditioning contrasts.** These paired 95% cluster-aware intervals correspond to the PSD and Fixed-PSD scores in Table 4. Both tracks compare the conditional PSD readout with its learned fixed counterpart.

Track	PSD − Fixed-PSD	Paired 95% interval
Remote	$+0.0064$	$[0.002, 0.010]$
Neighborhood	$+0.0066$	$[0.0023, 0.0109]$

Table 25: **External Signed retrieval metrics.** The same Signed and Fixed-Signed readouts as Table 4. Secondary metrics are descriptive.

Method	nDCG@10	Hit@1	MRR	AP
Remote				
Fixed-Signed	0.6090	0.6775	0.7386	0.5479
ReaPO-Signed	0.6399	0.6976	0.7551	0.5884
Neighborhood				
Fixed-Signed	0.7494	0.8022	0.8434	0.6596
ReaPO-Signed	0.7693	0.8218	0.8566	0.6862

Table 26: **External Signed conditioning gains across clusterings.** Signed-minus-Fixed-Signed nDCG@10 differences with nominal 95% intervals from 20,000 paired bootstrap draws under each cluster definition.

Contrast	Δ	Sequence component	CATH superfamily	Ligand identity
Remote				
S − Fixed-S	$+0.0309$	$[0.0246, 0.0378]$	$[0.0224, 0.0400]$	$[0.0234, 0.0399]$
Neighborhood				
S − Fixed-S	$+0.0199$	$[0.0124, 0.0293]$	$[0.0112, 0.0294]$	$[0.0138, 0.0276]$

Table 27: **External Signed conditioning gains by relation and seed.** Existing nDCG@10 aggregates; Δ is Signed minus Fixed-Signed. Relation rows average the three seeds; seed rows average the two relation axes with equal weight. Eligible folds and the three fixed support panels are averaged within each query–relation unit. These are two marginal summaries, not relation-by-seed cells or additional significance tests. Full-precision values appear in `protocol_supplement/external_axis_seed.csv`.

Track	Relation / training seed	Signed	Fixed-Signed	Δ
Remote	Structural superfamily	0.9283	0.8769	$+0.0514$
	Ligand identity	0.3516	0.3411	$+0.0105$
	20260806	0.6415	0.6056	$+0.0358$
	20260807	0.6394	0.6062	$+0.0332$
	20260808	0.6389	0.6152	$+0.0238$
Neighborhood	Structural superfamily	0.9642	0.9428	$+0.0214$
	Ligand identity	0.5744	0.5561	$+0.0183$
	20260806	0.7687	0.7482	$+0.0205$
	20260807	0.7704	0.7491	$+0.0213$
	20260808	0.7689	0.7510	$+0.0179$