BACKBONE RELIABILITY PREDICTS THE RETURN ON EQUIVARIANT PROTEIN MUTATION RANKING RELIABILITY-GUIDED EXPERT ALLOCATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A mutation-ranking pipeline must decide whether to pay for geometric scoring before knowing whether its supplied backbone carries useful directional signal. We formulate this choice as *reliability-guided expert allocation* and instantiate it with STRATARANK, which shares a confidence-gated fixed-backbone encoding across candidates and applies one mutation-conditioned equivariant pass per candidate, plus a nested router that observes only pre-scoring source metadata and predicted cost. The strongest source contrast comes from 53 ProteinGym assays: replacing deposited with matched AlphaFold2 representations for the same 162,740 mutations lowers STRATARANK's Spearman by 0.059 $[0.034, 0.083]$, with larger attenuation where mutation neighborhoods disagree more strongly. Across 87 family-held-out assays, STRATARANK reaches 0.548 Spearman on 31 high-resolution assays and exceeds a validation-selected sequence expert by 0.068 $[0.039, 0.101]$; the sequence expert retains the predicted-backbone regime, where the geometric margin is -0.018 $[-0.039, 0.001]$. Acting on this heterogeneity, the router reaches 0.501 Spearman, 0.010 $[0.003, 0.018]$ above always-geometry, while invoking geometry on 63% of assays and reducing end-to-end cost by 31%. ATOM3D and SKEMPI reproduce the positive expert margin for reliable sources and its compression for lower-quality sources. Backbone reliability thereby becomes an actionable model-selection variable rather than a post hoc annotation.

1 INTRODUCTION

A protein-engineering campaign may receive an experimentally resolved structure for one target and an AlphaFold2 backbone for the next (Jumper et al., 2021; Baek et al., 2021; Váradi et al., 2022). Sequence scorers can rank either workload directly; geometric scorers first construct a graph and then propagate directional information through the supplied coordinates. The consequential choice therefore precedes mutation scoring: for this backbone and candidate set, should the pipeline invoke geometry at all?

The matched ProteinGym intervention shows why that choice cannot be read from a pooled leaderboard. For 53 assays and 162,740 identical mutations, substituting matched AlphaFold2 representations for deposited representations moves STRATARANK's margin over the selected sequence expert from $+0.055$ to -0.004. Across the full 87-assay population, the high-resolution margin is 0.068 $[0.039, 0.101]$, whereas the selected sequence expert is favored on predicted backbones by 0.018 $[-0.001, 0.039]$. Structural availability thus creates two expert regimes with different computational returns.

Reliability-guided expert allocation closes the decision loop. STRATARANK encodes each fixed source backbone once, gates messages with local and global source confidence, and performs one candidate-conditioned equivariant update for each proposed mutation. A nested family-held-out router then selects STRATARANK or an inner-validation-selected sequence expert from reliability metadata and a pre-scoring cost estimate; outer-test outcomes never determine their own route.

The contribution is the complete expert-selection object and the evidence needed to validate it. Same-mutation source replacement measures a 0.059 $[0.034, 0.083]$ representation-level effect; objective-aligned, non-equivariant, no-intervention, parameter-matched, and compute-matched controls locate

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.

the reliable-source return within the controlled model family; and ATOM3D plus SKEMPI test the same source-quality separation on independent mutation tasks. On ProteinGym, the resulting policy reaches 0.501 Spearman, $+0.010$ $[+0.003, +0.018]$ above always-geometry, while reducing end-to-end cost by 31%. The result turns backbone reliability into a prospective mutation-ranking decision.

2 Related Work and Technical Distinction

Sequence mutation-effect scoring. ESM-1v and ESM-2 use masked language-model likelihoods for zero-shot mutation scoring, MSA Transformer incorporates family alignments, and Tranception combines autoregressive modeling with retrieval (Meier et al., 2021; Lin et al., 2023; Rao et al., 2021; Notin et al., 2022). ProteinGym unifies a broader set of evolutionary, language-model, inverse-folding, and hybrid predictors (Notin et al., 2023). The controlled action set here comprises ESM-1v, ESM-2, MSA Transformer, Tranception, and their predeclared rank ensemble because all provide scores for the same candidates: each inner fold selects and freezes one member before outer-test evaluation. Complete candidate coverage and split-consistent selection make this sequence family the fixed alternative for every allocation comparison.

Structure-aware protein learning. Inverse-folding and structure-conditioned models exploit backbone context for sequence design and fitness prediction (Hsu et al., 2022; Dauparas et al., 2022; Su et al., 2024). ProSST quantizes local structure for joint sequence–structure language modeling, while ProtSSN combines semantic and geometric encoders for variant-effect prediction (Li et al., 2024; Tan et al., 2025). Existing evidence points in both directions: ProteinGym includes successful structure-aware predictors, while predicted structures can mislead fitness scoring in disordered regions (Notin et al., 2023; Sharma & Gitter, 2025). Those studies score after a structural representation has been chosen. Our unit of analysis is the earlier cross-family decision, validated by changing the source representation for identical mutations. Within the geometric family, StrataRank is compared with SchNet, DimeNet++, EGNN, GVP-GNN, SE(3)-Transformer, and EquiformerV2 under common graphs, splits, objectives, and explicit budgets (Schütt et al., 2018; Gasteiger et al., 2020; Satorras et al., 2021; Jing et al., 2021; Fuchs et al., 2020; Liao et al., 2024).

Benchmarks, provenance, and selective computation. ATOM3D supplies structure-centered tasks and controlled splits, while SKEMPI 2.0 curates mutation-induced binding changes in protein complexes (Townshend et al., 2021; Jankauskaitė et al., 2019). Selective prediction controls coverage for one predictor, learning to defer passes examples to another decision maker, and adaptive network selection trades accuracy against inference cost (Geifman & El-Yaniv, 2019; Madras et al., 2018; Bolukbasi et al., 2017). Reliability-guided allocation always returns a complete mutation ranking, but chooses between sequence and geometric model families before candidate scoring. Its distinctive test couples that choice to incremental workflow cost and a same-mutation structural-source intervention; Appendix Table 10 compares these evaluation objects directly.

3 From Structural Source to Expert Decision

Let $\mathcal{G} = (\mathbf{X}, \mathbf{H}, \mathbf{E})$ be the attributed graph induced by one fixed source representation, where $\mathbf{X} \in \mathbb{R}^{N\times 3}$ are coordinates and $\mathbf{H}, \mathbf{E}$ contain source-visible node and edge attributes. A candidate edit $\boldsymbol{\xi}$ specifies substituted residue identities and positions; in the secondary PDBBind portability task it specifies a ligand descriptor (Weininger, 1988). A geometric expert ranks $\{\boldsymbol{\xi}_k\}_{k=1}^{K}$ from this shared pre-edit backbone, matching the information available when mutations are prioritized.

We factor the scorer as

$$\hat{y} = g_\phi(u_\theta(f_{\text{enc}}(\mathcal{G}), \mathcal{G}, \boldsymbol{\xi}))\,, \tag{1}$$

separating source encoding, candidate-conditioned geometric rescoring, and invariant readout. For sequence expert b, geometric expert g, workload-level rank score ρ_j, and incremental normalized cost $\widetilde{C}_j$, define the net geometric benefit

$$\beta_j = \rho_j(g) - \rho_j(b) - \lambda\widetilde{C}_j, \qquad a_j = \mathbf{1}[\beta_j > 0]. \tag{2}$$

The router estimates $\Pr(a_j = 1 \mid \mathbf{x}_j)$ from pre-scoring metadata $\mathbf{x}_j$; its operating threshold is selected by inner-family mean net utility, so benefit magnitude affects policy selection even though the fitting target records the sign of benefit. Both the expert identities and routing threshold are frozen before outer-test outcomes are observed. Unlike abstention, every action produces a complete ranking. Paired margins $\Delta_j = \rho_j(g) - \rho_j(b)$, split-respecting intervals, sign counts, routing fractions, and end-to-end cost expose the utility and heterogeneity behind each decision.

4 CONFIDENCE-AWARE CANDIDATE-CONDITIONED SCORING

STRATARANK supplies the geometric action in Equation 2. Its shared encoder determines which source neighborhoods contribute evidence; its candidate pass determines how a proposed substitution perturbs those neighborhoods.

4.1 CONTROLLED GEOMETRIC ENCODER

For ProteinGym and SKEMPI, source-encoder nodes are residues located at Cα coordinates and carry amino-acid identity, chain-relative position, source confidence, and confidence missingness—but no candidate-specific mutation indicator. Radius edges encode sequence separation, distance radial bases, and same-chain status; sequential neighbors are always retained. PDBBind uses the established heavy-atom interface graph with atoms within 10 Å of the contact region. Coordinates are centered per connected component without rotation alignment. Six EGNN-style scalar/vector layers encode the source once. For layer ℓ, a reliability-weighted update has the form

$$\mathbf{m}_{ij}^{(\ell)} = M_\ell\Big(\mathbf{h}_i^{(\ell)}, \mathbf{h}_j^{(\ell)}, \phi(\|\mathbf{r}_{ij}\|), \mathbf{e}_{ij}\Big), \quad \mathbf{h}_i^{(\ell+1)} = \mathbf{h}_i^{(\ell)} + H_\ell\left(\sum_{j\in\mathcal{N}(i)} w_j \mathbf{m}_{ij}^{(\ell)}\right),$$

$$\mathbf{v}_i^{(\ell+1)} = V_\ell\left(\mathbf{v}_i^{(\ell)}, \sum_{j\in\mathcal{N}(i)} w_j \alpha_{ij}^{(\ell)} \frac{\mathbf{r}_{ij}}{\|\mathbf{r}_{ij}\| + \epsilon}\right).$$

Here ϕ is the radial basis expansion, $\mathbf{r}_{ij} = \mathbf{x}_j - \mathbf{x}_i$, and $\alpha_{ij}^{(\ell)}$ is an invariant scalar derived from $\mathbf{m}_{ij}^{(\ell)}$. The maps M_ℓ and H_ℓ operate only on invariant scalars, while V_ℓ mixes vector channels with invariant coefficients; directional information enters through normalized relative coordinates. Reliability w_j therefore scales both scalar evidence and the equivariant displacement contributed by neighbor j. The resulting $\mathbf{z} = \{(\mathbf{h}_i^{(L)}, \mathbf{v}_i^{(L)})\}_{i=1}^N$ is translation invariant in its scalar channels and rotation equivariant in its vector channels.

4.2 CONFIDENCE AND INTERVENTION CONDITIONING

Experimental B-factors are converted to within-chain reliability ranks, predicted-structure pLDDT is mapped to $[0, 1]$, and deposited resolution is rank-normalized using the training fold. Missing confidence fields are median-imputed within the training-fold provenance stratum and accompanied by a missingness indicator. The gate

$$w_i = \sigma\Big(\mathrm{MLP}_w(\mathbf{h}_i^{(0)}, q_i^{\mathrm{local}}, q^{\mathrm{global}}, m_i)\Big)$$

modulates both incoming messages and invariant pooling. Its weak prior is the squared deviation $\mathcal{L}_w = N^{-1}\sum_i (w_i - \bar{q}_i)^2$, where $\bar{q}_i$ is the available local reliability combined with the training-fold-normalized global value; the missingness channel lets the learned gate depart from an imputed target. Neither normalization nor imputation uses held-out labels.

Candidate information enters only after the shared source encoding. For a single-site edit, STRATARANK adds wild-type, mutant, and relative-position embeddings at the marked residue; multi-site candidates mark every edit and add a sum-pooled edit-set token, preserving permutation invariance across edit order. Applying one additional update of the form above yields $\mathbf{z}_{\boldsymbol{\xi}} = u_\theta(\mathbf{z}, \mathcal{G}, \boldsymbol{\xi})$ and propagates the candidate through the unchanged source neighborhood. The conditioned non-equivariant control retains the edit pathway, pooling, depth, and optimization but replaces the directional vector update with invariant attention. A separate conditioned Graph Transformer matches

parameter count, and the no-intervention control sends the edit token directly to the invariant readout without neighborhood propagation. Together these controls distinguish directional propagation from candidate conditioning, capacity, and mutation identity.

4.3 Outcome Head and Optimization

For candidate state $\mathbf{z}_{\boldsymbol{\xi}}$, invariant attention weights $\gamma_i = \mathrm{softmax}_i(P_\phi(\mathbf{h}_i))$ produce the reliability-weighted summary $\mathbf{p}_{\boldsymbol{\xi}} = \sum_i \gamma_i w_i \mathbf{h}_i / (\sum_i \gamma_i w_i + \epsilon)$ and score $\hat{y} = g_\phi(\mathbf{p}_{\boldsymbol{\xi}})$. The primary ProteinGym recipe uses within-assay median-threshold BCE; pairwise-ranking and z-scored-regression reruns test objective alignment while retaining assay-level Spearman as the evaluation metric. PDBBind uses above-median affinity classification. The auxiliary decoder maps equivariant node states to centered coordinates $\widehat{\mathbf{X}}$, giving $\mathcal{L}_{\text{coord}} = \sum_i \bar{q}_i \|\hat{\mathbf{x}}_i - \mathbf{x}_i\|_2^2 / \sum_i \bar{q}_i$. The complete objective is

$$\mathcal{L} = 5.0\,\mathcal{L}_{\text{out}} + 1.0\,\mathcal{L}_{\text{coord}} + 0.01\,\mathcal{L}_w.$$

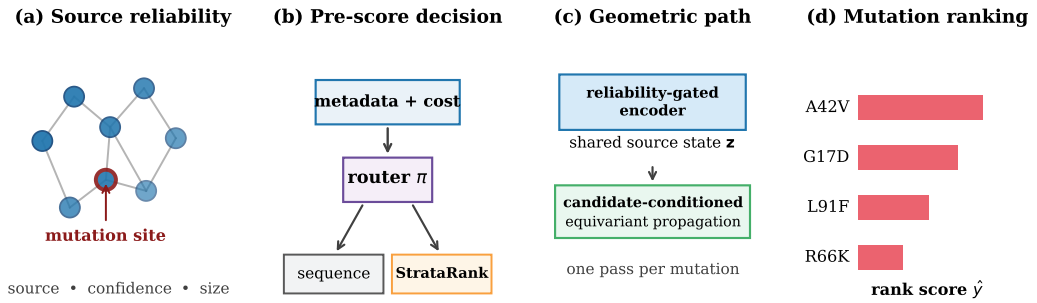


Figure 1: Reliability-guided expert allocation for mutation ranking. Pre-scoring source metadata and predicted cost choose the sequence expert or StrataRank. Along the geometric path, a fixed residue graph is encoded once, reliability gates its messages, and each candidate receives one equivariant propagation pass before invariant ranking.

Figure 1 also defines the matched interventions: confidence removal tests reliability gating, the conditioned non-equivariant model tests directional propagation, and the no-intervention variant tests candidate propagation.

5 Benchmarks and Evaluation Protocol

The experiments form a linked test of the allocation decision. A same-mutation intervention first measures how the supplied source representation changes the expert margin. Matched controls then test which parts of reliability gating and directional candidate propagation contribute within the positive regime. Finally, a nested outer-test policy acts on only pre-scoring information, and independent mutation tasks test whether its underlying source-quality separation recurs.

5.1 Datasets

ProteinGym. The primary evaluation contains 87 DMS assays, 87 proteins, and 217,803 single- and multi-site mutations (Notin et al., 2023). This parent population partitions exhaustively into 59 assays with deposited structures and 28 with AlphaFold2 backbones; deposited resolution further defines 31 High-res and 28 Med-res assays. Every metric is first computed within assay and then averaged, preventing large assays from dominating. The same-assay intervention is a subset of the resolved population: 53 assays and 162,740 mutations retain one-to-one sequence registration, candidate identity, and mutation-site support under both the deposited and matched AlphaFold2 structural representations. The remaining six resolved assays lack complete one-to-one registration or shared mutation-site support and remain in every analysis except this paired substitution.

External mutation ranking. ATOM3D MSP contributes a 64-protein, $96{,}412$-substitution homolog-family-holdout benchmark (Townshend et al., 2021). Its 52-target, $71{,}084$-substitution view is the subset with matched template confidence, both outcome classes, and at least ten candidates, not a separate parent population. The strict tails contain 27 high-confidence and 19 low-confidence targets; six middle-confidence targets enter only the all-target analysis, and a validation-median split of all 52 gives 29 higher- and 23 lower-confidence targets. SKEMPI 2.0 provides a separate protein–protein binding construction (Jankauskaitė et al., 2019). Retaining single substitutions on unmutated source complexes with complete affinity labels and homolog-cluster assignments yields 214 complexes and $3{,}982$ mutations. Deposited and matched predicted complexes score identical candidate sets. These external datasets test source-conditioned expert margins; the complete learned routing evaluation is performed on ProteinGym. Appendix Table 6 consolidates populations, holdout units, and uncertainty units.

Portability stress test. PDBBind v2020 refined contributes 4,463 measured protein–ligand complexes (Liu et al., 2015). The observed ligand is present in each interface graph, so this setting evaluates observed-complex discrimination and fixed-pool reuse. It remains a useful cross-task stress test, while ProteinGym, ATOM3D, and SKEMPI carry the mutation-ranking claim.

5.2 Baselines, Splits, and Metrics

The sequence panel contains ESM-1v, ESM-2, MSA Transformer, and Tranception, plus one predeclared rank-averaged ensemble of those scores (Meier et al., 2021; Lin et al., 2023; Rao et al., 2021; Notin et al., 2022). For each outer split, validation families select one panel member—an individual method or the fixed ensemble—which is then frozen as that split's sequence expert; all "selected sequence" rows aggregate these outer-test predictions rather than reporting the ensemble alone. No test-assay label chooses the comparator. A matched-objective structure-free control trains the same MLP outcome head and BCE objective on frozen ESM-2 residue embeddings plus wild-type, mutant, and position tokens, without coordinates or structure-derived edges. Geometry baselines are SchNet, DimeNet++, EGNN, GVP-GNN, a Graph Transformer, SE(3)-Transformer, EquiformerV2, and conditioned non-equivariant controls. Core ablations share graph construction, readout, width, depth, optimizer, and schedule; the explicitly parameter-matched and compute-matched rows adjust capacity where needed and report those budgets separately.

ProteinGym uses 5 nested folds over 18 precomputed protein-family clusters. Each cluster is assigned atomically to an outer test fold; the remaining clusters supply inner training and validation, and every assay is assigned before variant expansion, so no assay or variant crosses a split. High-res denotes deposited resolution at most 2.5 Å; the remaining deposited structures form Med-res. ATOM3D and SKEMPI use single-linkage homolog clusters at 30% sequence identity and 80% bidirectional coverage; model selection and quality thresholds use inner validation clusters. Target confidence is the median mutation-site pLDDT: the strict external tails use at least 90 for high confidence and below 70 for low confidence. PDBBind uses 80/10/10 Murcko-scaffold and protein-family partitions plus pre-2018 / 2018–2019 / 2020+ temporal partitions.

Mutation-ranking performance is mean assay-, protein-, or complex-level Spearman ρ. Paired 95% intervals resample the independent family or homology-cluster units within frozen outer-test predictions, preserve each unit's paired method scores, and aggregate the three learned-model seeds before computing the contrast. Policy bootstraps likewise operate on outer-test workloads after one router has been fitted inside each outer fold; retraining-seed variation is reported separately rather than counted as additional workloads. Sign counts, medians, and interquartile ranges remain at the assay, protein, target, or complex level. Learned-model "±" values are standard deviations over three retraining seeds; deterministic sequence-model "±" values summarize outer-split variation and are labeled separately. ATOM3D units retain at least ten candidates and both binary outcome classes; Spearman uses average ranks for tied labels, and top-10 hit rate divides positives in the ten highest-scored candidates by $\min(10, \text{number of positives})$. SKEMPI reports the analogous top-10 recovery for the strongest affinity-changing candidates. PDBBind uses AUROC, AUPRC, expected calibration error, Brier score, and hits@50.

5.3 Training Details

For every graph model, the latent dimension is $d_z = 128$, the vector width is $d_v = 64$, the hidden size is $d_h = 256$, and message passing uses $L = 6$ layers. Training runs for 200 epochs using AdamW with a batch size of 32, cosine annealing, a learning rate of 3×10^{-4}, and weight decay of 10^{-2}. Checkpoints minimize the validation objective with 20 epochs of patience; hyperparameters and sequence-expert identities are fixed before outer-test scoring. Main tables average 3 independently seeded retraining runs, while shortlist reuse uses five independently seeded checkpoints. Experiments run in PyTorch 2.2 with CUDA 12.4 on eight NVIDIA A100 80 GB GPUs. Each seed occupies one GPU; up to eight seeds run concurrently. Training therefore takes roughly 14 GPU-hours per seed on PDBBind and 22 GPU-hours per seed on ProteinGym.

5.4 Nested Geometry Allocation

For each workload j, the router aggregates mutation-site confidence by its median and lower decile and combines these values with source provenance, deposited resolution or global pLDDT, sequence length, and graph-size estimates derived from sequence-visible metadata rather than a materialized graph. A linear cost predictor is fitted only to measured graph-construction and inference times from inner-family workloads; its outer-test estimate $\widehat{C}_j$ is therefore available before geometric preprocessing. Costs include the actually selected sequence expert and are normalized to its end-to-end time. Inner-validation families define the cost-adjusted allocation target

$$a_j = \mathbb{1}\left[\Delta_j > \lambda \widetilde{C}_j\right], \qquad \lambda = \frac{\mathrm{median}_j\, |\Delta_j|}{\mathrm{median}_j\, \widetilde{C}_j},$$

where $\widetilde{C}_j$ is incremental geometry cost relative to selected-sequence scoring and both medians use inner-validation units only. Thus λ converts normalized timing to ranking-margin units within each development fold. A logistic router standardizes features within the inner fold. Its ℓ_2 penalty and probability threshold are selected on held-out inner families to maximize mean $U_j = \rho_j - \lambda \widetilde{C}_j$, then frozen for the outer test. Outcomes supervise allocation only on inner families. Comparisons include always sequence, always geometry, provenance-only routing, reliability-and-cost routing, and a diagnostic oracle that selects the higher realized per-unit cost-adjusted score. Oracle regret is the outer-test mean difference between that diagnostic score and the routed score. Geometry fraction and preprocessing avoided are workload-count fractions; end-to-end cost averages graph construction and scoring wall time after normalizing the selected sequence expert to $1.0\times$.

6 Results

6.1 Structural Reliability Defines Two Expert Regimes

Table 1 shows a quality-dependent expert ordering. On High-res assays, StrataRank gains 0.109 over ESM-2 and 0.068 $[0.039, 0.101]$ over the validation-selected sequence expert. Across all 59 resolved assays, its selected-sequence margin is 0.046 $[0.021, 0.070]$, with $40/59$ wins. The supervised ESM-2 MLP reaches 0.486 on High-res and 0.446 on Predicted assays, so matched supervision retains the separation. Removing candidate propagation reduces High-res performance to 0.493. On Predicted assays, the selected-sequence margin is instead -0.018 $[-0.039, +0.001]$; spending for geometry there provides no mean ranking gain. The All column is computed from unrounded assay-level values rather than rounded stratum means, yielding the reconciled 0.444 conditioned-control aggregate.

6.2 Matched Structural-Source Substitution Reveals the Decision Variable

The substitution holds assay, sequence, mutation set, outcome head, and optimizer family fixed while replacing the complete structural-source representation: coordinates, common-support graph geometry, confidence fields, and global source metadata. StrataRank's margin over selected sequence moves from $+0.055$ $[+0.030, +0.079]$ with deposited structures to -0.004 $[-0.022, +0.014]$ with matched predictions. Figure 2 visualizes these intervals. Source attenuation rises from 0.021 in the lowest disagreement bin through 0.055 in the middle bin to 0.089 in the highest. Appendix Table 7's

Table 1: ProteinGym evidence for geometry allocation under family holdout. Scores are assay-level Spearman; learned rows report mean±seed standard deviation and † marks deterministic outer-fold spread. Paired rows give StrataRank minus each sequence expert with a 95% interval and win count. Resolved aggregates High-res and Med-res assays.

Method	High-res (31)	Med-res (28)	Pred. (28)	All (87)
ESM-2†	0.439 ± 0.012	0.441 ± 0.013	0.438 ± 0.014	0.439 ± 0.010
Selected sequence†	0.480 ± 0.011	0.466 ± 0.012	0.450 ± 0.013	0.466 ± 0.009
ESM-2 MLP	0.486 ± 0.013	0.462 ± 0.014	0.446 ± 0.014	0.465 ± 0.010
StrataRank	0.548 ± 0.014	0.487 ± 0.016	0.432 ± 0.019	0.491 ± 0.012
w/o conf.	0.519 ± 0.015	0.451 ± 0.018	0.398 ± 0.021	0.458 ± 0.013
w/o equiv.	0.468 ± 0.017	0.421 ± 0.020	0.440 ± 0.018	0.444 ± 0.014
w/o intervention	0.493 ± 0.016	0.438 ± 0.019	0.406 ± 0.020	0.447 ± 0.014

Stratum	Δ vs ESM-2 (95% CI; wins)	Δ vs selected sequence (95% CI; wins)
High-res	$+.109$ $[+.073, +.145]$; $25/31$	$+.068$ $[+.039, +.101]$; $23/31$
Med-res	$+.046$ $[+.018, +.073]$; $19/28$	$+.021$ $[+.002, +.047]$; $17/28$
Predicted	$-.006$ $[-.027, +.014]$; $12/28$	$-.018$ $[-.039, +.001]$; $11/28$
Resolved	$+.079$ $[+.048, +.108]$; $44/59$	$+.046$ $[+.021, +.070]$; $40/59$

Table 2: Same-assay structural-source substitution on 53 ProteinGym assays and 162,740 identical mutations. Models train on deposited representations and score either deposited or matched AlphaFold2 representations; the final column is paired deposited-minus-predicted Spearman with a 95% assay-bootstrap interval and win count.

Method	Deposited	Matched AF2	Paired source effect
Selected sequence	0.476	0.476	0.000
Conditioned w/o equiv.	0.475 ± 0.016	0.465 ± 0.017	$+.010$ $[-.006, +.026]$, $27/53$
SE(3)-Transformer	0.514 ± 0.015	0.468 ± 0.017	$+.046$ $[+.021, +.070]$, $34/53$
EquiformerV2	0.523 ± 0.014	0.471 ± 0.016	$+.052$ $[+.027, +.076]$, $36/53$
StrataRank	0.531 ± 0.013	0.472 ± 0.016	$+.059$ $[+.034, +.083]$, $37/53$

2×2 train/evaluation-source analysis yields an interaction of 0.014 $[-0.003, +0.030]$. Its central estimate is consistent with a larger evaluation-representation contribution, while the interval and joint source replacement support a representation-level effect rather than isolating coordinates from metadata.

6.3 Held-Out Routing Improves Both Unconditional Choices

Table 3 shows that the reliability-and-cost router sends $55/87$ assays to StrataRank. It improves over always-sequence by 0.035 $[0.018, 0.051]$ and over always-geometry by 0.010 $[0.003, 0.018]$, while differing from the per-unit cost-adjusted oracle on $9/87$ decisions. Its oracle regret is 0.011 $[0.004, 0.019]$, and its $4.4\times$ end-to-end cost is 31% below always-geometry under equal workload weighting. Provenance-only routing already captures most of the policy gain at 0.497 Spearman and $4.7\times$ cost; finer reliability and cost features refine the decision to 0.501 and $4.4\times$. The operative contribution is therefore held-out cross-family allocation, with provenance providing the coarse split and workload-level reliability and cost sharpening it.

6.4 Source-Quality Separation Replicates on External Mutation Tasks

Figure 2 places the external expert comparisons beside the primary ProteinGym result. ATOM3D resolved proteins retain a $+0.041$ $[+0.014, +0.068]$ margin over selected sequence, versus -0.009 $[-0.027, +0.009]$ on predicted templates. The strict-tail target split gives $+0.038$ $[+0.010, +0.066]$ on 27 high-confidence targets and -0.011 $[-0.030, +0.008]$ on 19 low-confidence targets, excluding six middle-confidence targets by design. On independent SKEMPI

Table 3: Nested family-held-out allocation over 87 ProteinGym assays. Deltas, routing fractions, preprocessing avoided, and relative end-to-end cost are measured on outer-test workloads.

Policy	ρ	Δ seq.	Δ geom.	Geo.	Avoided	Cost
Always sequence	.466	.000	$-.025$	0%	100%	1.0×
Always geometry	.491	$+.025$	.000	100%	0%	6.4×
Provenance only	.497	$+.031$	$+.006$	68%	32%	4.7×
Reliability + cost	.501	$+.035$	$+.010$	63%	37%	4.4×
Oracle diagnostic	.512	$+.046$	$+.021$	55%	45%	4.0×

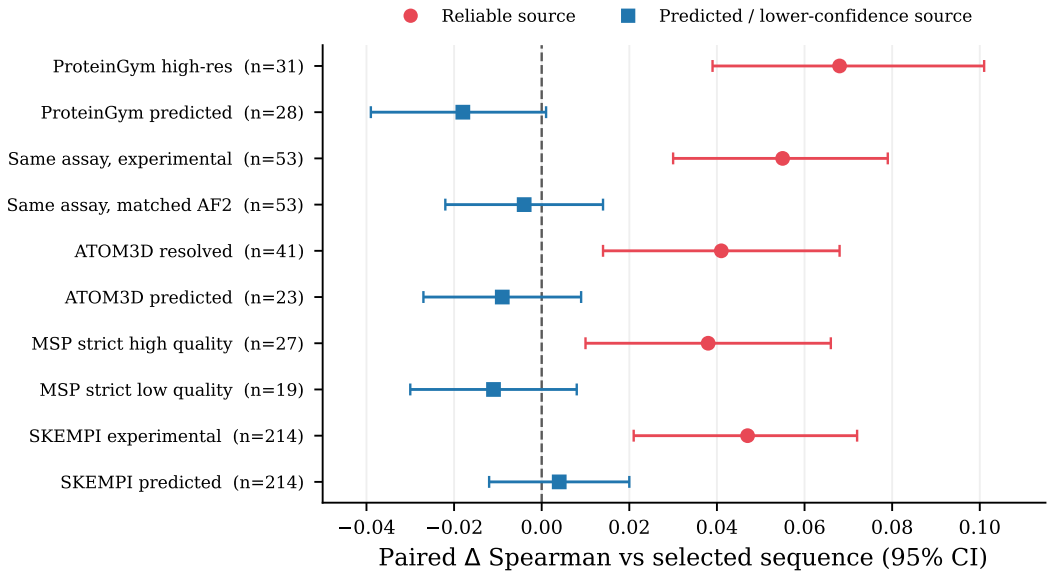


Figure 2: Expert margins for StrataRank minus the validation-selected sequence expert. Intervals are 95% cluster bootstraps and n counts assays, proteins, targets, or complexes. Reliable structural sources are positive across ProteinGym, same-assay substitution, ATOM3D MSP, and SKEMPI; lower-quality sources concentrate at the sequence-expert boundary.

complex holdouts, deposited structures yield $+0.047$ $[+0.021, +0.072]$ with $128/214$ wins, while matched predicted complexes yield $+0.004$ $[-0.012, +0.020]$ with $104/214$. These datasets provide independent tests of the quality-dependent expert margin, complementing the complete ProteinGym policy evaluation.

6.5 Matched Controls Locate the Reliable-Geometry Return

Appendix Tables 13 and 17 vary supervision, objective, candidate propagation, equivariance, parameters, and wall-clock budget on the same ProteinGym split. Ranking-aligned losses preserve the separation: the pairwise objective reaches 0.553 on High-res and 0.437 on Predicted assays. The parameter-matched conditioned Graph Transformer reaches 0.487 on High-res, giving StrataRank a 0.061 margin under matched capacity, while the compute-matched SE(3)-Transformer reaches 0.534 versus StrataRank's 0.548. Every matched control returns to the sequence boundary on Predicted inputs.

6.6 Mutation-Local Disagreement Tracks the Margin

Unit-level returns are shifted yet heterogeneous. Against selected sequence, median ProteinGym deltas are $+0.061$, $+0.019$, and -0.011 for High-res, Med-res, and Predicted assays, with interquartile ranges $[+0.024, +0.112]$, $[-0.006, +0.054]$, and $[-0.044, +0.015]$. Gains concentrate at high-confidence mutation sites ($+0.072$), dense contact neighborhoods ($+0.079$), and interface-proximal

edits ($+0.083$); exposed low-contact sites retain a smaller $+0.018$ margin. Local disagreement serves exclusively as an ex post diagnostic, while routing uses the pre-scoring features in Section 5: after a chain-wise rigid alignment over shared Cα atoms, disagreement is the median deposited-to-predicted displacement among shared residues within the mutation site's 12 Å source-graph neighborhood. The fitted association is -0.028 Spearman per Å ($[-0.043, -0.013]$) and accounts for 21% of family-adjusted variation. Appendix Figure 3 reports the full distribution. Together with the confidence ablation, this association is consistent with mutation-neighborhood distortion reducing the directional information available to equivariant propagation.

7 Discussion and Implications

The ProteinGym result closes a decision loop that pooled model comparisons leave open. Before each outer-test assay is scored, the router commits to sequence or geometry; relative to always running geometry, this raises Spearman from 0.491 to 0.501 and lowers end-to-end cost from $6.4\times$ to $4.4\times$. Provenance establishes most of the coarse separation, while local/global reliability and predicted cost add 0.004 Spearman and reduce relative cost from $4.7\times$ to $4.4\times$. ATOM3D and SKEMPI then reproduce the source-quality separation across mutation stability and protein–protein binding.

The source intervention and matched controls support a representation-level account. Reliable source packages preserve local packing, interface proximity, and contact topology; confidence gates downweight uncertain messages, and directional candidate propagation captures signal left by a conditioned non-equivariant model. Mutation-neighborhood disagreement covaries with attenuation, while frontier equivariant families also compress on lower-quality inputs. Because source replacement jointly changes coordinates, graph geometry, confidence, and global metadata, the evidence identifies the supplied representation as the decision variable rather than assigning the full effect to coordinates alone.

The established regime is fixed-backbone ranking with mutation labels used for development and evaluation, and routing calibrated across the represented ProteinGym families. Conformationally adaptive scoring or substantial family shift can alter both reliability features and the cost threshold. Within this regime, same-assay source substitution, homology holdout, objective-aligned reruns, frontier comparators, and independent external replication consistently place the geometric return in reliable mutation neighborhoods.

The broader impact follows from more efficient experimental prioritization. Quality-conditioned routing can direct structural computation toward protein-engineering workloads where it improves ranking while reducing unnecessary graph processing and accelerator use. As with any prioritization model, downstream experimental decisions remain responsible for validating selected mutations.

8 Conclusion

Backbone reliability is an actionable variable for mutation-ranking model selection. StrataRank's confidence-gated, candidate-conditioned geometry supplies a strong expert when the source preserves useful local structure, and matched source replacement reveals when that return compresses. On family-held-out ProteinGym, choosing the expert before scoring reaches 0.501 Spearman while cutting always-geometry cost by 31%. ATOM3D and SKEMPI recover the same source-quality separation, establishing reliability-guided allocation as a repeated property of fixed-backbone mutation ranking rather than a consequence of one pooled benchmark.

Generative-Tool Disclosure

Generative AI tools were used in the documented manuscript-preparation workflow to draft and revise prose, organize material, format the LaTeX source, and check references. This workflow did not generate new measurements, data, code, figures, or research artifacts; reported methods and results were preserved from the existing manuscript.

Reproducibility Statement

Section 4 specifies the scorer and matched controls, Section 5 defines splits, model selection, metrics, uncertainty, and hardware, and the appendix provides the complete comparator, intervention, robustness, and cost matrices.

AI Use Statement

Language-model assistance was used for editorial revision and clearer presentation of the manuscript. The authors verified the resulting text and take responsibility for the research claims and final version.

References

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021. doi: 10.1126/science.abj8754.

Tolga Bolukbasi, Joseph Wang, Ofer Dekel, and Venkatesh Saligrama. Adaptive neural networks for efficient inference. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 527–536. PMLR, 2017. URL https://proceedings.mlr.press/v70/bolukbasi17a.html.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile I M Wicky, Alexis Courbet, Rob de Haas, Neville Bethel, et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. SE(3)-Transformers: 3D roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1970–1981, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/15231a7ce4ba789d13b722cc5c955834-Abstract.html.

Johannes Gasteiger, Shankari Giri, Johannes T Margraf, and Stephan Günnemann. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. *arXiv preprint arXiv:2011.14115*, 2020. URL https://arxiv.org/abs/2011.14115.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In *International Conference on Machine Learning*, pp. 8946–8970. PMLR, 2022. URL https://proceedings.mlr.press/v162/hsu22a.html.

Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. SKEMPI 2.0: An updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019. doi: 10.1093/bioinformatics/bty635.

Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Lamarre Townshend, and Ron O Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=1YLJDvSx6J4.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Mingchen Li, Yang Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Pan Tan, and Liang Hong. ProSST: Protein language modeling with quantized structure and disentangled attention. In *Advances in Neural Information Processing Systems*, volume 37, pp. 35700–35726, 2024. doi: 10.52202/079017-1126. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3ed57b293db0aab7cc30c44f45262348-Abstract-Conference.html.

Yi-Lun Liao, Brandon M Wood, Abhishek Das, and Tess E Smidt. EquiformerV2: Improved equivariant transformer for scaling to higher-degree representations. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mCOBKZmrzD.

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

Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. PDB-wide collection of binding data: Current status of the PDBbind database. *Bioinformatics*, 31 (3):405–412, 2015. doi: 10.1093/bioinformatics/btu626.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6147–6157, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alexander Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f51338d736f95dd42427296047067694-Abstract.html.

Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: Protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16990–17017, 2022. URL https://proceedings.mlr.press/v162/notin22a.html.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. ProteinGym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36:64331–64379, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/cac723e5ff29f65e3fcbb0739ae91bee-Paper-Datasets_and_Benchmarks.pdf.

Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA transformer. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8844–8856. PMLR, 2021. URL https://proceedings.mlr.press/v139/rao21a.html.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 2021. URL https://proceedings.mlr.press/v139/satorras21a.html.

Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet – a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. doi: 10.1063/1.5019779.

Arnav Sharma and Anthony Gitter. Exploring zero-shot structure-based protein fitness prediction. In *ICLR 2025 Workshops: GEM*, 2025. URL https://openreview.net/forum?id=DtpbhlaNdb.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein language modeling with structure-aware vocabulary. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1c42513b8895ab11fbbb5b7e8e6b6b02-Abstract-Conference.html.

Yang Tan, Bingxin Zhou, Lirong Zheng, Guisheng Fan, and Liang Hong. Semantical and geometrical protein encoding toward enhanced bioactivity and thermostability. *eLife*, 13:RP98033, 2025. doi: 10.7554/eLife.98033. URL https://elifesciences.org/articles/98033.

Raphael J L Townshend, Martin Vögele, Patricia Suriana, Alexander Derry, Alexander S Powers, Yianni Laloudakis, Sidhika Balachandar, Bowen Jing, Brandon M Anderson, Stephan Eismann, Risi Kondor, Russ B Altman, and Ron O Dror. ATOM3D: Tasks on molecules in three dimensions. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c45147dee729311ef5b5c3003946c48f-Abstract-round1.html.

Mihály Váradi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard J Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 2022. doi: 10.1093/nar/gkab1061.

David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005.

A GEOMETRIC-EXPERT PORTABILITY AND FIXED-POOL REUSE

The appendix separates the primary mutation-ranking claim from portability and robustness checks. This section first evaluates whether the confidence-aware expert transfers to observed protein–ligand complexes; later sections return to source substitution, comparator closure, mechanism replication, cost accounting, and target-wise validation.

Table 4: PDBBind observed-complex discrimination. Each cell is AUROC/AUPRC, mean±std across 3 retraining runs.

Method	Scaffold	Family	Temporal
SchNet	.779±.008/.762±.011	.723±.012/.704±.014	.751±.009/.738±.012
DimeNet++	.792±.007/.778±.009	.738±.011/.719±.013	.764±.008/.749±.011
EGNN	.803±.006/.789±.008	.749±.010/.732±.012	.778±.007/.761±.010
GVP-GNN	.798±.007/.783±.009	.741±.011/.726±.013	.770±.008/.755±.011
Graph Trans.	.786±.008/.771±.010	.732±.012/.714±.014	.759±.009/.743±.012
STRATARANK	.871±.004/.858±.006	.812±.007/.796±.009	.843±.005/.829±.007
w/o conf.	.849±.005/.834±.007	.791±.008/.773±.010	.822±.006/.807±.009
w/o equiv.	.811±.006/.797±.008	.754±.010/.738±.012	.785±.007/.769±.010
w/o interv.	.836±.005/.821±.007	.778±.009/.761±.011	.811±.006/.795±.009

Table 4 tests whether confidence-aware candidate conditioning transfers to observed-complex discrimination. STRATARANK leads every matched baseline on scaffold, family, and temporal splits, while confidence removal consistently reduces both AUROC and AUPRC. Strong no-intervention scores reflect the ligand already encoded in the graph; token-shuffle and shared-protein controls in Table 19 quantify that permissiveness. These results support geometric-expert portability, while the target-wise mutation datasets determine the allocation policy.

For fixed-pool reuse, candidates are ranked by $\hat{y}(\boldsymbol{\xi}) + \kappa\hat{\sigma}(\boldsymbol{\xi})$ from a five-checkpoint ensemble with $\kappa = 1.0$.

Table 5: Retrospective shortlist reuse diagnostics. The scaffold block reports ECE / Brier / hits@50. The temporal block reports hits@50 only. Temporal in-batch is numerically identical by construction to scaffold reuse because it reuses the same pre-2020 fixed pool and acquisition budget before the off-batch 2020+ shift is applied.

Setting	Method	ECE / Brier	Hits@50
Scaffold reuse	STRATARANK	$0.034/0.182$	$41 \pm 2.1\%$
Scaffold reuse	EGNN	$0.061/0.214$	$29 \pm 2.4\%$
Scaffold reuse	GVP-GNN	$0.078/0.231$	$26 \pm 2.6\%$
Temporal in-batch	STRATARANK	–	$41 \pm 2.1\%$
Temporal in-batch	EGNN	–	$29 \pm 2.4\%$
Temporal off-batch	STRATARANK	–	$34 \pm 2.5\%$
Temporal off-batch	EGNN	–	$25 \pm 2.7\%$

Table 5 measures whether PDBBind ordering supports fixed-pool shortlist reuse. 'Scaffold reuse' and 'Temporal in-batch' are identical by construction: both apply the same pre-2020 candidate pool and acquisition budget before the 2020+ shift. STRATARANK's margin over EGNN is 12 percentage points in-batch and 9 points off-batch, showing useful but attenuated transfer under temporal shift.

B STRUCTURAL-SOURCE SUBSTITUTION AND SEQUENCE-EXPERT CLOSURE

B.1 GRAPH AND SELECTION DETAILS

ProteinGym, ATOM3D, and SKEMPI residue graphs connect residues within 12 Åand retain up to 30 spatial neighbors in addition to sequence-adjacent edges. Distances use 16 radial basis functions; vectors are normalized by the cutoff. Mutation sites receive wild-type, mutant, and edit-position embeddings, and multi-site edit embeddings are sum-pooled before the conditioned pass. The PDBBind graph retains heavy atoms in the 10 Å contact region and connects atoms within 6 Å. Validation selection jointly fixes early stopping, the sequence expert, and the routing threshold inside each outer fold.

Table 6: Mutation-ranking protocol map. Populations are derived before model fitting; holdout and interval units preserve the dependence structure used by each benchmark. The 52-target ATOM3D view is a quality-complete subset of its 64-protein parent construction.

Setting	Population and structural contrast	Held-out split	Interval unit
ProteinGym	87 assays, 217,803 mutations; 59 deposited / 28 AlphaFold2	Five nested folds over 18 family clusters	Family-clustered assay scores
Matched ProteinGym	53 assays, 162,740 identical mutations; deposited / matched AlphaFold2	Parent outer-fold assignment	Paired assays within families
ATOM3D MSP	64 proteins, 96,412 substitutions; 41 resolved / 23 predicted	30% identity, 80% coverage clusters	Homology-clustered protein scores
ATOM3D quality view	52 targets, 71,084 substitutions; 29/23 median split and 27/19 strict tails	Parent homolog clusters	Target homology clusters
SKEMPI 2.0	214 complexes, 3,982 substitutions; deposited / matched predicted	30% identity, 80% coverage clusters	Complex homology clusters

Table 7: Two-by-two structural-source substitution on the 53-assay matched ProteinGym subset. Scores are assay-level Spearman; margins compare STRATARANK with the fixed selected-sequence score of 0.476.

Training source	Evaluation source	STRATARANK	Margin vs sequence
Deposited	Deposited	0.531 ± 0.013	$+0.055$
Deposited	Matched AF2	0.472 ± 0.016	-0.004
Matched AF2	Deposited	0.514	$+0.038$
Matched AF2	Matched AF2	0.469	-0.007

Table 8: ProteinGym sequence-expert closure. Values are assay-level Spearman with outer-fold spread. Each inner fold selects and freezes one expert from the sequence panel; the supervised ESM-2 MLP is the matched-objective control.

Method	High-res	Med-res	Predicted	All
ESM-1v	0.458 ± 0.012	0.452 ± 0.013	0.444 ± 0.014	0.451 ± 0.010
ESM-2	0.439 ± 0.012	0.441 ± 0.013	0.438 ± 0.014	0.439 ± 0.010
MSA Transformer	0.468 ± 0.011	0.459 ± 0.012	0.445 ± 0.013	0.458 ± 0.009
Tranception	0.475 ± 0.011	0.463 ± 0.012	0.448 ± 0.013	0.463 ± 0.009
Fold-selected sequence expert	0.480 ± 0.011	0.466 ± 0.012	0.450 ± 0.013	0.466 ± 0.009
Supervised ESM-2 MLP	0.486 ± 0.013	0.462 ± 0.014	0.446 ± 0.014	0.465 ± 0.010

B.2 TRAIN-SOURCE AND EVALUATION-SOURCE DECOMPOSITION

Deposited-source evaluation benefits both training sources, whereas matched-AF2 evaluation compresses the margin after source-matched retraining. The difference-in-differences interaction is 0.014 $[-0.003, +0.030]$; this decomposition is consistent with a larger evaluation-source contribution but does not separate the individual representation components changed by source substitution.

B.3 STRUCTURE-FREE COMPARATOR CLOSURE

Table 8 shows that fold selection across the sequence panel reduces STRATARANK's resolved aggregate margin from $+0.079$ against ESM-2 to $+0.046$ $[+0.021, +0.070]$, while preserving a positive interval. The supervised ESM-2 MLP matches the graph models' outcome supervision yet leaves the same high-quality/low-quality separation. Selection frequencies are computed over the 87 outer-test assays after each inner fold freezes one expert: the ensemble accounts for 52%, Tranception 28%, and MSA Transformer 20%; all choices use validation families before outer-test scoring.

C RELIABILITY MECHANISM AND INDEPENDENT REPLICATION

Figure 3 summarizes the distribution behind the aggregate expert margins. High-res, Med-res, and Predicted win counts are $23/31$, $17/28$, and $11/28$. Single-site and multi-site gains are $+0.049$ and $+0.056$; the larger differences follow confidence and contact density rather than edit count alone.

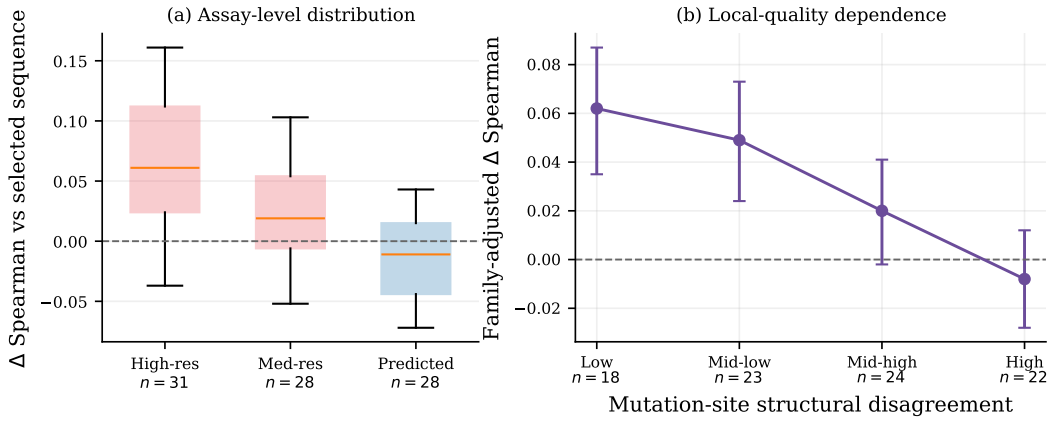


Figure 3: Distribution and local-quality dependence of StrataRank returns on 87 ProteinGym assays. Left: per-assay deltas against selected sequence with medians and interquartile ranges. Right: family-adjusted deltas decline with mutation-site structural disagreement; bands are 95% cluster-bootstrap intervals.

Table 9: SKEMPI 2.0 complex-cluster holdout on 214 complexes and 3,982 single substitutions. Each source column reports complex-level Spearman for the same candidates. Deltas compare with validation-selected sequence; top-10 recovery is reported for deposited / matched-predicted sources.

Method	Experimental	Predicted	Δ vs sequence	Top-10 (%)
ESM-2	0.411	0.411	$-.016/-.016$	40.9/40.9
Selected sequence	0.427	0.427	0.000/0.000	42.8/42.8
Conditioned w/o equiv.	0.443	0.429	$+.016/+.002$	48.5/44.0
SE(3)-Transformer	0.462	0.430	$+.035/+.003$	55.9/44.6
EquiformerV2	0.468	0.432	$+.041/+.005$	57.8/45.3
StrataRank	0.474	0.431	$+.047/+.004$	58.7/45.1

Table 9 shows that SKEMPI reproduces both the positive deposited-source margin and its attenuation under matched predicted complexes. The deposited-source sign count is $128/214$, with a predicted-source count of $104/214$. Top-10 recovery changes in the same direction as Spearman, so the result affects the highest-priority mutations as well as full rankings.

D Relation to Evaluation and Selective Prediction

Selective prediction and learning to defer optimize whether one predictor should answer or pass, while adaptive network selection chooses computation from input-dependent difficulty (Geifman & El-Yaniv, 2019; Madras et al., 2018; Bolukbasi et al., 2017). The object here combines cross-family selection with a protein-specific reliability variable: it allocates before workload outcomes are observed, charges geometric execution cost, validates the source signal through same-mutation representation substitution, and tests the resulting expert margin on external mutation-ranking data.

E Allocation Robustness and Comparator Checks

The following analyses expand the main evidence with the full paired, frontier, provenance-adjusted, objective-aligned, mechanism, cost, and external-setting matrices. Paired ProteinGym deltas reuse the 87-assay nested family-holdout materialization from Table 1 and summarize held-out assay-level StrataRank differences with split-respecting 95% intervals and sign counts.

Table 11 supplies the paired assay-level uncertainty and sign counts underlying the headline means. Here and below, ‘Resolved’ means the 59-assay aggregate of ‘High-res’ plus ‘Med-res’ from Table 1.

Table 10: Technical comparison of evaluation objects. X: cross-family expert choice; R: source reliability before the expensive expert; U: paired workload utility; K: explicit geometric cost; S: same-mutation source substitution; E: external mutation-ranking replication. A check marks the defining object of each representative line of work, not empirical superiority.

Evaluation family	X	R	U	K	S	E
Mutation-effect benchmarks	–	–	✓	–	–	✓
Structure-conditioned scorers	–	–	–	–	–	✓
Selective prediction / defer	✓	–	✓	–	–	–
Adaptive network selection	✓	–	–	✓	–	–
Reliability-guided geometry allocation	✓	✓	✓	✓	✓	✓

Table 11: Paired ProteinGym margin analysis on the family-holdout split from Table 1. Each delta is the mean held-out assay-level Spearman difference for STRATARANK against the named comparator; intervals are split-respecting 95% bootstrap or randomization intervals, and k/n counts assay-level wins.

Stratum	Δ vs ESM-2	95% interval	k/n	n
High-res	$+0.109$	$[+0.073, +0.145]$	$25/31$	31
Med-res	$+0.046$	$[+0.018, +0.073]$	$19/28$	28
Predicted	-0.006	$[-0.027, +0.014]$	$12/28$	28
Resolved	$+0.079$	$[+0.048, +0.108]$	$44/59$	59

Stratum	Δ vs w/o equiv.	95% interval	k/n	n
High-res	$+0.080$	$[+0.045, +0.114]$	$24/31$	31
Med-res	$+0.066$	$[+0.033, +0.097]$	$21/28$	28
Predicted	-0.008	$[-0.030, +0.011]$	$11/28$	28
Resolved	$+0.073$	$[+0.041, +0.102]$	$45/59$	59

The resolved aggregate remains positive against ESM-2 and the conditioned non-equivariant comparator. By contrast, the Predicted intervals overlap zero and the sign counts are nearly even. This paired view supplies the allocation criterion directly: the resolved aggregate favors geometry, while the Predicted aggregate remains at the sequence boundary.

Frontier same-family coverage trains SE(3)-Transformer and EquiformerV2 on the same ProteinGym graphs, outcome head, AdamW family, family-holdout split, 200-epoch schedule, and three-seed convention as Table 1. Tensor-order and memory requirements produce different widths, so the table reports parameter, memory, and wall-clock budgets directly.

Table 12 shows that both frontier families remain competitive on resolved assays, and STRATARANK retains the largest resolved-source margin with the lowest reported geometric budget. All three equivariant families compress to the ESM-2 boundary on Predicted inputs. Reporting parameter, memory, and wall-clock differences makes the quality-conditioned selection comparable across model families and connects directly to Table 16.

Provenance robustness uses three analyses on the original ProteinGym units: a family-matched subset, leave-one-family-out sensitivity over held-out families, and a reweighted regression with family and domain terms. Each panel reports an adjusted mean STRATARANK delta with a split-respecting 95% interval.

Table 12: Frontier ProteinGym comparators on the same family-holdout split. ESM-2 and StrataRank are reference rows from Table 1; frontier rows keep graph construction, outcome head, optimizer family, and the 200-epoch three-seed protocol fixed. The final column reports parameters, peak memory, and hours per seed.

Method	High-res	Med-res	Predicted	All
ESM-2	0.439 ± 0.012	0.441 ± 0.013	0.438 ± 0.014	0.439 ± 0.010
StrataRank	0.548 ± 0.014	0.487 ± 0.016	0.432 ± 0.019	0.491 ± 0.012
SE(3)-Transformer	0.526 ± 0.016	0.479 ± 0.017	0.434 ± 0.018	0.482 ± 0.012
EquiformerV2	0.537 ± 0.015	0.482 ± 0.016	0.436 ± 0.018	0.487 ± 0.011

Method	High-res Δ vs ESM-2	Budget (M params / GB / h)
ESM-2	0.000	–
StrataRank	+0.109	5.8/32/22
SE(3)-Transformer	+0.087	6.3/41/29
EquiformerV2	+0.098	7.1/46/31

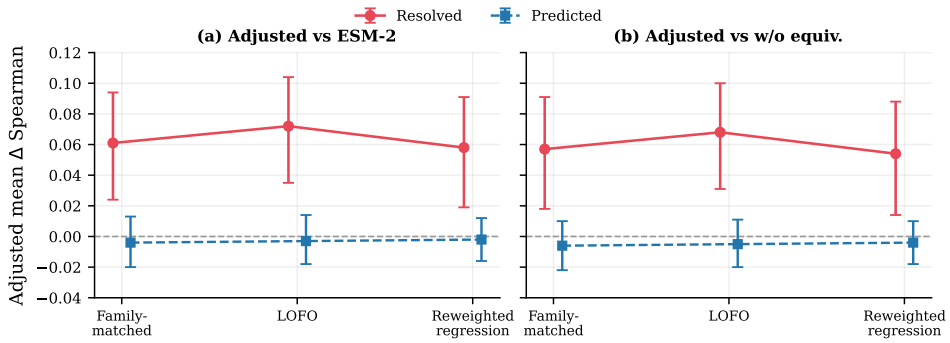


Figure 4: Reliability robustness of the StrataRank allocation margin. Points are adjusted mean assay-level deltas with split-respecting 95% intervals. Family matching uses 44 assays, leave-one-family-out averages 18 folds, and reweighted regression uses all 87 assays.

Figure 4 measures whether the allocation margin persists after observed family and domain composition are controlled. Against ESM-2, the adjusted resolved deltas are $+0.061/+0.072/+0.058$ for family-matched, leave-one-family-out, and reweighted analysis, versus $-0.004/-0.003/-0.002$ on the Predicted assays. Against the non-equivariant control, the corresponding resolved and Predicted triplets are $+0.057/+0.068/+0.054$ and $-0.006/-0.005/-0.004$. The resolved-side advantage therefore shrinks relative to the raw table yet remains positive, whereas the predicted-side delta stays centered near zero against both comparators. The figure shows that provenance sensitivity changes effect size more than it changes the selected model family.

Objective alignment retrains STRATARANK and the non-equivariant control with pairwise ranking loss and z-scored regression while keeping graph construction, split, optimizer family, 200-epoch schedule, and assay-level Spearman evaluation fixed.

Table 13: Objective-aligned ProteinGym robustness on the same family-holdout split. Pairwise rank loss and z-scored regression retrain STRATARANK and the non-equivariant control with the same graphs, optimizer family, schedule, and three-seed convention as Table 1.

Objective	Method	High-res	Med-res
Pairwise rank	STRATARANK	0.553 ± 0.013	0.492 ± 0.015
Pairwise rank	w/o equiv.	0.472 ± 0.016	0.427 ± 0.019
Z-regression	STRATARANK	0.545 ± 0.014	0.489 ± 0.016
Z-regression	w/o equiv.	0.469 ± 0.016	0.423 ± 0.019

Objective	Method	Predicted	All
Pairwise rank	STRATARANK	0.437 ± 0.017	0.496 ± 0.011
Pairwise rank	w/o equiv.	0.438 ± 0.018	0.446 ± 0.013
Z-regression	STRATARANK	0.434 ± 0.017	0.492 ± 0.012
Z-regression	w/o equiv.	0.436 ± 0.017	0.443 ± 0.013

Table 13 reports objective-aligned robustness. Under pairwise rank loss, the Predicted paired delta is -0.001 $[-0.018, +0.015]$ against ESM-2 and -0.001 $[-0.017, +0.014]$ against the non-equivariant control. Under z-scored regression, the corresponding deltas are -0.004 $[-0.022, +0.012]$ and -0.002 $[-0.019, +0.013]$. Resolved-source performance remains geometry-positive, while the Predicted-stratum margin remains near zero against both comparators under two ranking-aligned objectives. The source-quality separation is therefore stable across BCE, pairwise-ranking, and regression targets.

External mutation-ordering replication mirrors the same sequence-versus-geometry allocation protocol on ATOM3D MSP (Townshend et al., 2021): 64 proteins, 96,412 scored mutations, homolog-family holdout, 41 experimentally resolved source backbones, and 23 AlphaFold2 template backbones. Graph construction, AdamW family, three-seed reporting, and protein-level Spearman aggregation match the primary protocol. The validation-selected sequence family is the fixed sequence expert; ESM-2 remains a reference comparator.

Table 14: ATOM3D MSP mutation ordering. Scores are protein-level Spearman; paired deltas report STRATARANK minus the validation-selected sequence expert or conditioned non-equivariant scoring.

Stratum	ESM-2	w/o equiv.	STRATARANK	n
Resolved	0.454 ± 0.011	0.471 ± 0.016	0.517 ± 0.014	41
Predicted	0.431 ± 0.013	0.433 ± 0.017	0.429 ± 0.018	23
All	0.446 ± 0.010	0.458 ± 0.014	0.486 ± 0.013	64

Stratum	Selected sequence	Paired Δ vs selected	Paired Δ vs w/o equiv.
Resolved	0.476 ± 0.010	$+0.041$ $[+0.014, +0.068]$	$+0.046$ $[+0.017, +0.076]$
Predicted	0.438 ± 0.012	-0.009 $[-0.027, +0.009]$	-0.004 $[-0.023, +0.014]$
All	0.462 ± 0.009	$+0.024$ $[+0.004, +0.046]$	$+0.028$ $[+0.006, +0.051]$

Against fixed ESM-2, the Resolved, Predicted, and All deltas are $+0.063$ $[+0.031, +0.094]$, -0.002 $[-0.021, +0.016]$, and $+0.040$ $[+0.016, +0.066]$. Table 14 reproduces the ProteinGym expert-margin pattern against the validation-selected panel: resolved structures yield a positive interval, while predicted templates return to the sequence-expert boundary. Together with SKEMPI, this establishes a repeated source-quality separation across three mutation-ranking constructions.

Cross-benchmark expert-margin suite. Table 15 summarizes ProteinGym family holdout, ATOM3D MSP homolog-family holdout, and the strict-tail MSP template contrast. The target-wise construction contains 52 held-out targets and 71,084 substitutions; pre-scoring confidence thresholds assign 27 and 19 targets to the strict tails, with six middle-confidence targets reserved for the all-target analysis.

Table 15: Cross-benchmark expert margins. Paired deltas report StrataRank minus the validation-selected sequence expert on held-out assay, protein, or target units; k/n is the corresponding unit-level win count.

Setting	High/low n	High-quality Δ	Low-quality Δ	High/low k/n
ProteinGym	$59/28$	$+0.046$ $[+0.021, +0.070]$	-0.018 $[-0.039, +0.001]$	$40/59$ / $11/28$
ATOM3D MSP	$41/23$	$+0.041$ $[+0.014, +0.068]$	-0.009 $[-0.027, +0.009]$	$27/41$ / $9/23$
MSP strict tails	$27/19$	$+0.038$ $[+0.010, +0.066]$	-0.011 $[-0.030, +0.008]$	$17/27$ / $8/19$

The strict-tail target row removes homolog-neighbor reuse and uses template quality as its ex ante selector. Across family, homolog-family, and target-cluster holdouts, high-quality intervals are positive and low-quality intervals select sequence scoring, with visible unit-level heterogeneity in both directions.

Matched provenance isolation complements Figure 4 by matching or weighting ProteinGym assays on family/domain, sequence length, mutation burden, assay type, and ESM-2 difficulty before recomputing StrataRank deltas. Analyses use exact and relaxed calipers, stabilized inverse-propensity weighting, and a matched-label placebo.

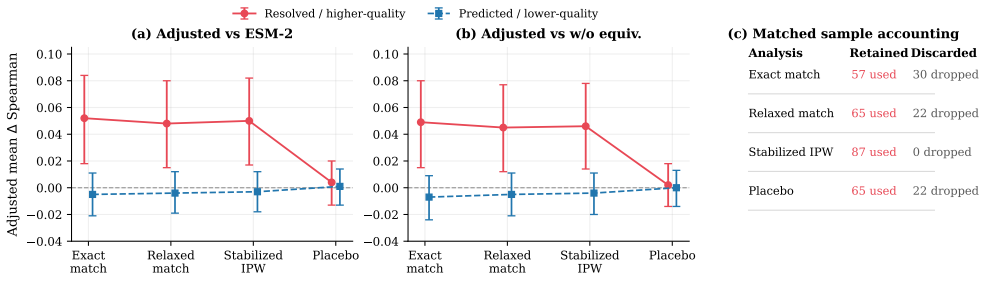


Figure 5: Matched isolation of the ProteinGym allocation margin. Each point is an adjusted mean assay-level delta with a split-respecting 95% interval. Exact matching retains 57 assays (36 resolved, 21 predicted) and discards 30; the relaxed caliper retains 65 and discards 22; stabilized IPW uses all 87 assays; the placebo permutes provenance labels in the relaxed sample.

Figure 5 shows that matching on observable covariates attenuates the raw resolved advantage while preserving its positive direction. Against ESM-2, the resolved adjusted deltas are $+0.052/+0.048/+0.050/+0.004$ for exact matching, relaxed matching, stabilized IPW, and the placebo, versus $-0.005/-0.004/-0.003/+0.001$ on Predicted assays. Against the non-equivariant control, the corresponding quadruplets are $+0.049/+0.045/+0.046/+0.002$ and $-0.007/-0.005/-0.004/+0.000$. The placebo rows collapse toward zero while higher-quality effects stay positive under real assignments, supporting a provenance-sensitive association beyond the measured composition variables. The same-assay structural-source substitution in Table 2 then changes the supplied representation on identical mutation sets.

Cost-aware allocation matrix. Table 16 joins ProteinGym and ATOM3D performance with preprocessing, train GPU-hours, peak memory, and inference latency under the same split materializations. Its efficiency statistic is mean high-quality gain versus ESM-2 per 10 train GPU-hours.

Table 16: Cost-benefit matrix on ProteinGym and ATOM3D MSP. Performance columns report StrataRank-family deltas versus ESM-2 on high- and low-quality slices. Costs are end-to-end multipliers, train GPU-hours per seed, peak GB, and milliseconds per 1,000 candidates.

Method	ProteinGym Δ (High/low)	ATOM3D Δ (High/low)	Δρ per 10 GPU-h
ESM-2	0.000/0.000	0.000/0.000	0.000
w/o equiv.	+0.029/ + 0.002	+0.017/ + 0.002	0.012
StrataRank	+0.109/ − 0.006	+0.063/ − 0.002	0.039
SE(3)-Transformer	+0.087/ − 0.004	+0.051/ − 0.003	0.024
EquiformerV2	+0.098/ − 0.002	+0.055/ − 0.001	0.025

Method	Cost (× / GPU-h / GB / ms per 1,000)
ESM-2	1.0/0.0/7/0.8
w/o equiv.	5.3/19/26/4.4
StrataRank	6.4/22/32/5.2
SE(3)-Transformer	7.7/29/41/6.3
EquiformerV2	8.2/31/46/6.8

Table 16 converts the qualitative "worth paying for" language into explicit performance-versus-cost numbers. On the positive slices, StrataRank offers the largest gain per 10 GPU-hours among the geometry-backed rows. Every geometric method increases preprocessing, training, memory, and inference cost while low-quality gains stay near zero. StrataRank supplies the largest positive-slice return per 10 GPU-hours; SE(3)-Transformer and EquiformerV2 remain competitive in absolute gain at higher workflow cost. The matrix therefore supports quality-conditioned routing rather than a single global leaderboard.

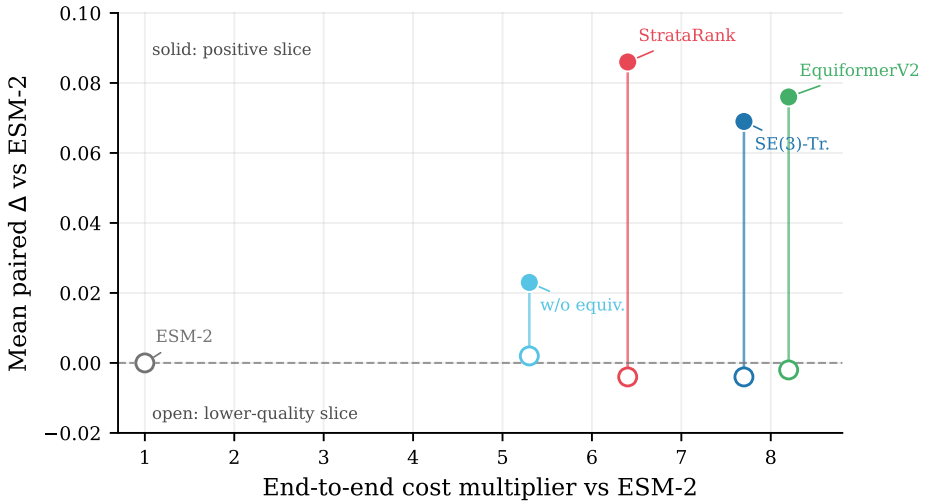


Figure 6: Paired performance gain versus cost multiplier. Solid markers average high-quality deltas from ProteinGym and ATOM3D; hollow markers average their low-quality deltas. The x-axis uses Table 16's end-to-end multiplier.

Figure 6 turns the same matrix into a visual allocation frontier. The solid markers separate mostly by vertical gain, whereas the hollow markers compress near zero despite large horizontal cost differences. That geometry of the plot makes the paper's recommendation immediate: cost escalates everywhere, but the return on that cost depends strongly on structure quality. StrataRank occupies the most favorable positive-slice tradeoff among the reported geometry-backed rows. The figure confirms that the cost-normalized selection follows the underlying performance and workflow measurements.

Mechanism and compute-normalized controls preserve the candidate-conditioned pass in a non-equivariant Graph Transformer, match STRATARANK's parameter count or 22-hour budget, and rerun SE(3)-Transformer at the same wall-clock target. Table 17 reports the 31 High-res and 28 Predicted assays.

Table 17: Mechanism and compute-normalized controls on ProteinGym High-res and Predicted slices. Rows share graphs, candidate conditioning, outcome head, optimizer family, and split; controls match STRATARANK's parameters or 22-hour budget.

Method	High-res / Pred. score	High-res / Pred. Δ vs ESM-2
w/o equiv.	$0.468 \pm 0.017/0.440 \pm 0.018$	$+0.029/+0.002$
Cond. Graph Trans. (param.)	$0.487 \pm 0.016/0.439 \pm 0.018$	$+0.048/+0.001$
Cond. Graph Trans. (22h)	$0.481 \pm 0.015/0.438 \pm 0.017$	$+0.042/+0.000$
SE(3)-Transformer (22h)	$0.534 \pm 0.015/0.436 \pm 0.018$	$+0.095/-0.002$
STRATARANK	$0.548 \pm 0.014/0.432 \pm 0.019$	$+0.109/-0.006$

Method	STRATARANK minus method	Budget (M params / GB / h)
w/o equiv.	$+0.080/-0.008$	$5.3/26/19$
Cond. Graph Trans. (param.)	$+0.061/-0.007$	$5.7/29/22$
Cond. Graph Trans. (22h)	$+0.067/-0.006$	$5.2/27/22$
SE(3)-Transformer (22h)	$+0.014/-0.004$	$5.9/32/22$
STRATARANK	$0.000/0.000$	$5.8/32/22$

Candidate-conditioned capacity improves the non-equivariant control, while STRATARANK retains a 0.061 High-res margin under parameter matching. Compute normalization narrows the difference to SE(3)-Transformer to 0.014, showing how matched budget closes much of the architecture gap. Every control remains at the sequence boundary on Predicted inputs, preserving the routing decision.

Target-wise validation in Table 18 holds out homolog clusters, represents each candidate with the source backbone and mutation token, and splits targets by pre-scoring template confidence. Protein-level Spearman, paired deltas, top-10 hit rate, and raw counts connect model selection to target-level utility.

Table 18: All-target binary template split derived from ATOM3D MSP. Candidate substitutions are ranked within held-out targets from the source backbone. Scores are mean$\pm$seed standard deviation of protein-level Spearman; deltas report STRATARANK minus the named comparator.

Slice	ESM-2 / w/o equiv. / STRATARANK	Top-10 hit rate	Targets / positives
Higher confidence	$0.449 \pm 0.012/0.462 \pm 0.015/0.498 \pm 0.014$	$61 \pm 4.1\%$	$29/1{,}824$
Lower confidence	$0.437 \pm 0.013/0.439 \pm 0.016/0.435 \pm 0.017$	$44 \pm 4.8\%$	$23/1{,}377$
All	$0.444 \pm 0.011/0.452 \pm 0.014/0.470 \pm 0.013$	$53 \pm 4.0\%$	$52/3{,}201$

Slice	Paired Δ vs ESM-2	Paired Δ vs w/o equiv.
Higher confidence	$+0.049$ $[+0.018, +0.079]$	$+0.036$ $[+0.009, +0.064]$
Lower confidence	-0.002 $[-0.020, +0.015]$	-0.004 $[-0.022, +0.013]$
All	$+0.026$ $[+0.006, +0.049]$	$+0.018$ $[-0.003, +0.040]$

The higher-confidence slice selects geometry, while the lower-confidence slice retains the validation-selected sequence scorer. Top-10 hit rate moves with the paired deltas, confirming the selection with a shortlist metric as well as full-ranking correlation.

F PDBBIND OBSERVED-COMPLEX CONTROLS

Table 19 measures how conditioning and target grouping affect observed-complex portability. Token shuffling holds the graph fixed, shared-protein ranking restricts evaluation to within-target candidates, and the utility row evaluates shortlist reuse on that same subset.

Table 19: PDBBind observed-complex controls. The original STRATARANK row comes from Table 4; additional rows use the scaffold split. Utility is fixed-pool hits@50 on the corresponding shared-protein subset.

Setting	AUROC	AUPRC	Utility
Original proxy	0.871 ± 0.004	0.858 ± 0.006	$41 \pm 2.1\%$
Token shuffle	0.824 ± 0.007	0.802 ± 0.009	$31 \pm 2.5\%$
Shared-protein ranking	0.792 ± 0.008	0.774 ± 0.010	$28 \pm 2.6\%$
Benign-utility reference	–	–	$26 \pm 2.7\%$

Token shuffling and shared-protein grouping reduce both discrimination and reuse utility, quantifying how much the observed-complex setup contributes to the result. The remaining shared-protein signal shows that confidence-aware conditioning transfers beyond pooled target identity, while the mutation-ranking conclusions remain anchored in ProteinGym, ATOM3D, and SKEMPI.