A permutation-null artifact drives a conservation–divergence spectrum's signature trend, and phylogeny dominates a protein language model's per-position coupling signal

Huazhang Shen

Independent Researcher · hzshen@gmail.com · ORCID 0009-0006-8208-3111

Preprint, not peer reviewed. Version date: 2026-08-22. Part of a three-paper companion series on the conservation–divergence spectrum (Research Square, 2026).

Abstract

This is one of three companion preprints: P1 introduces the conservation–divergence spectrum, P2 maps when it can be trusted, and this paper (P3) cross-examines what the spectrum and a black-box protein language model each know about the same positions. A conservation–divergence spectrum's signature negative trend between conservation and divergence — the pattern the companion preprints read as evidence of specificity encoding — is *mostly the shape of its own permutation null* (observed ρ ≈ −0.61 here versus null ρ = −0.90; the real signal is the excess above that null, not the trend itself), and this null shape recurs across four independent families (GPCR, opsin, nuclear-receptor DBD, HLA), explaining ≥100% of the observed trend in three of four. Independently, a protein language model's (PLM) per-position signal at the same positions is dominated by phylogeny rather than mechanism: about three-quarters of its apparent knowledge (median clade-aware balanced accuracy 0.41 vs random-split 0.66; phylogeny 74%, pure coupling 24% above chance) reflects evolutionary relatedness, not the functional label it is being asked to predict — a clean quantitative case for the "PLMs are homology shortcuts" critique. We established both results by cross-examining the spectrum's per-position specificity score against a light probe on ESM-2 embeddings, position by position, on one test case: G-protein-coupling specificity across 219 class A GPCRs, 212 structurally equivalent positions, three coupling classes (Gs / Gi-o / Gq-11). We report six results. (1) The spectrum's own two internal scores correlate at ρ = 0.88, but the spectrum and the PLM probe correlate at only ρ = 0.17 — they rank different positions as important. (2) The signature negative trend is mostly the null's shape, as above. (3) That null shape recurs across four families — in three of four the null explains ≥100% of the observed trend, while HLA is a weaker case (78%) — the mechanistic root of the companion P2's finding that divergence thresholds cannot transfer between families. (4) Phylogeny accounts for about three-quarters of the PLM's per-position signal, as above. (5) Where the spectrum and the PLM probe disagree, the PLM probe (layer 30) is significantly enriched on the true structural coupling interface (Mann–Whitney p = 0.003) while the spectrum is not — but this is a distribution-level shift, not a top-list win (Fisher n.s.), and the interface is only a *partial* gold standard, so the spectrum's non-enrichment is not evidence it is wrong. (6) The conclusions are robust across ESM layers and probe types, with one claim actively falsified: an apparent bias of the pure-coupling residual toward conserved positions appears only at layer 30 and is an artifact

of that layer. As a negative control we add opsin spectral tuning, a family the companion P2 established as a miss: there **both** methods fail to recover the tuning code (top-list Fisher n.s. for each) and disagree with each other ($\rho = 0.005$) — the expected signature when the biology is genuinely unreadable. The contribution is a reusable framework for interrogating a black-box model's per-position knowledge and cross-checking it against independent evolutionary evidence, with the null-model discipline that makes such a cross-examination trustworthy.

Keywords: specificity-determining positions; protein language model; ESM-2; phylogenetic confound; null model; GPCR coupling; cross-method validation

Author summary

A headline pattern used in our companion preprints to flag specificity-determining positions turns out to be mostly a statistical artifact of the test itself — visible only once we compare against a properly shuffled null — and this caution applies to the whole field of sequence-based specificity detection. We show this, and a second, independent finding, by putting two very different tools side by side on one well-studied problem: how G-protein-coupled receptors choose their downstream partner. One tool, the conservation–divergence spectrum of our companion preprints, compares amino-acid identities across functional groups; it is transparent but only sees specificity written in individual residues. The other, a protein language model, has read millions of sequences and encodes each position in a rich numerical vector that can capture subtle, distributed patterns — but nobody can easily say what it has actually learned, or whether its apparent knowledge is real biology or just a reflection of which proteins are evolutionary cousins. Position by position, we ask where the two agree, where they disagree, and who is right when they do. Three-quarters of what the language model seems to know at each position turns out to be explained by evolutionary relatedness, not mechanism. Where the two methods disagree, the language model aligns better with the physically measured coupling interface, but weakly, and only as a shift in a distribution rather than a clean list of winners. A family where the biology is known to be unreadable serves as a control: there both methods fail together, exactly as they should. The framework, and the null-model discipline behind it, transfer to any attempt to audit what a black-box sequence model knows.

Introduction

Identifying the sequence positions that distinguish functional subtypes of a protein family — specificity-determining positions — is a decades-old problem with two currently active families of solution. The first is evolutionary and per-position: mutual-information partitions, cumulative relative-entropy scores, evolutionary trace, statistical coupling analysis, and the conservation–divergence spectrum of the companion preprints (P1, P2) all read, in different ways, the signature of subtype-specific selection at individual alignment columns. These methods are interpretable — each score belongs to a named residue — but they share a blind spot: by construction they detect specificity only where it is written in amino-acid identity at single positions, and they are systematically blind to specificity encoded in a distributed, epistatic, context-dependent way.

The second family is the protein language model. Trained by masked-language modeling on tens of millions of sequences, models such as ESM-2 [2] embed each position as a high-dimensional vector that folds in the whole-sequence context around it. That context is exactly what per-position evolutionary methods discard, so a PLM can in principle carry the distributed signal they miss. But a PLM is a black box: its embeddings are unlabeled, and a persistent, unresolved critique holds that much of what a PLM appears to "know" at a position is not mechanism but a homology shortcut [3,4] — a reflection of the training distribution's phylogenetic structure rather than of function. This paper subtracts that phylogenetic component explicitly and reports what remains.

These two approaches are rarely brought face to face at the resolution where they actually differ — the individual position. Receptor-level coupling prediction from PLM embeddings is well established (PRECOGx [1] predicts GPCR G-protein and β-arrestin coupling from ESM1b embeddings), but combining a protein language model with per-position resolution, an independent evolutionary cross-check, and explicit phylogenetic subtraction is, to our knowledge, not something existing GPCR-coupling work has done together.

This paper is that cross-examination. We treat the conservation–divergence spectrum and an ESM-2 probe as **two detectives on one case**: a *physical-evidence* detective that reads per-position evolutionary evidence, and an *intuition* detective that reads the PLM's contextual fingerprint. The metaphor is not decoration — it fixes the paper's question. We do not ask which detective scores higher on a benchmark; we ask where they agree, where they disagree, who aligns with independent physical truth at the points of disagreement, and — throughout — what null-model and phylogenetic discipline is needed before either detective's testimony can be believed. (This "two detectives" framing is a division of *methods*; it is distinct from the companion P2's two *dimensions* of applicability, which classify when a family is readable at all.) The test case is class A GPCR G-protein coupling: 219 receptors, 212 structurally equivalent generic-numbered positions, three coupling classes. A family the companion P2 established as unreadable — opsin spectral tuning — serves as a negative control.

Results

The test case and the three scores

All three scores are read from the same matrix: 219 class A GPCRs aligned on the GPCRdb generic-numbering frame, 212 positions present across the panel, each receptor labeled by primary coupling class (Gs, Gi-o, or Gq-11). The three scores differ only in *how* they read that matrix, so every comparison below is apples-to-apples — same receptors, same labels, same positions.

- **Physical-evidence, S_sdp.** The companion P1's headline per-position specificity-anomaly score: between-class Jensen–Shannon divergence minus the value predicted by a linear fit of divergence on within-class conservation across the family — i.e., the residual above the family-wide conservation–divergence trend (P1's anomaly_score; see P1 Methods). This is distinct from P1's auxiliary sdp_score (a conservation × divergence product used for P1's own ranking

and bootstrap-stability checks), which correlates with anomaly_score at only ρ = 0.78 across the same 212 positions and is not the score carried into this paper.

- **Physical-evidence, net, S_net.** The same evidence, but with the permutation-null floor subtracted per position (Results §"The signature trend is mostly the null"), expressed as a Gamma-fit tail probability so that positions are comparable across families. S_net is the evidence score we carry into cross-family comparison.

- **Intuition, S_plm.** A logistic-regression probe on ESM-2 per-position embeddings (esm2_t30_150M_UR50D, layer 21 for the main result), scored by balanced accuracy at predicting coupling class. The crucial point: **ESM-2 knows nothing about coupling** — it was trained only on masked amino-acid prediction and emits an unlabeled 640-dimensional context fingerprint; the coupling question is injected entirely by the probe's labels. After phylogenetic correction (clade-aware cross-validation), S_plm is the "pure coupling" signal.

The six results below split into two tracks that answer different questions. Three concern **what the two detectives actually find** — where S_sdp/S_net and S_plm agree and disagree (below), what the intuition detective's disagreements track structurally, and how robust that finding is. The other three concern **null discipline** — the discipline this companion paper's title refers to as "cross-examining": how much of the raw conservation–divergence trend is the null's shape rather than signal, whether that null shape is family-specific or universal, and how much of the PLM channel is phylogeny rather than biology. The negative control that closes the section tests both tracks together on a family where the biology itself is unreadable.

The two detectives barely agree

Comparing the three scores position by position (Fig 1) separates a within-method baseline from the cross-method result. The two physical-evidence scores, being two readings of the same evolutionary quantity, correlate strongly: S_sdp versus S_net, Spearman ρ = 0.88. But physical evidence versus intuition — S_net versus S_plm — correlate at only ρ = 0.17. The two detectives, given the same case, largely nominate *different* positions as important.

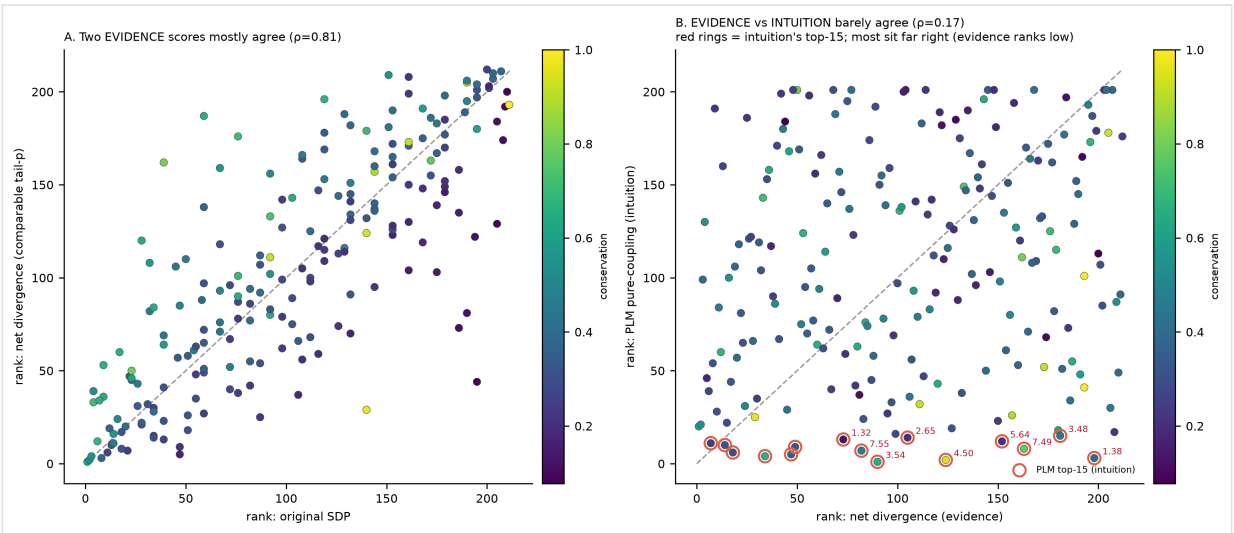

Fig 1. Three per-position scores on 212 GPCR positions. The two physical-evidence readings agree (S_sdp vs S_net, ρ = 0.88); physical evidence versus PLM intuition barely agree (S_net vs S_plm, ρ = 0.17). Disagreement is directional, not noise (see gold-standard anchoring, Fig 5).

This near-independence is the paper's starting fact. It could be noise — two weak detectors disagreeing at random — or it could be directional, the two methods reading genuinely different layers of the signal. The gold-standard anchoring below shows it is directional.

The signature trend is mostly the shape of the null

The conservation–divergence spectrum's signature observation, across the companion preprints, is a negative relationship between the two axes: more conserved positions are less between-class divergent (here, Spearman ρ ≈ −0.61). Taken at face value this looks like a biological law. It is mostly a statistical one.

For each position we built a permutation null by shuffling the coupling labels while preserving amino-acid composition, and recomputed the divergence. The null trend between conservation and divergence is itself strongly negative — **ρ_null = −0.90**, stronger than the observed −0.61 (Fig 2). The mechanism is a floor that rises as conservation falls: at highly conserved positions the null divergence is near zero, while at variable positions sampling noise inflates it, manufacturing a negative slope with no functional content. The signal that matters is therefore the **excess above the null**, not the trend — and 118 of 212 positions carry divergence significantly above their null ($z > 2$). Framed this way the signature trend is a control to be subtracted, not a result; and once subtracted, the conserved-yet-divergent corner that P1 reads as the specificity anchor is exactly where the null floor is near zero, so the anchor is reinforced rather than undermined.

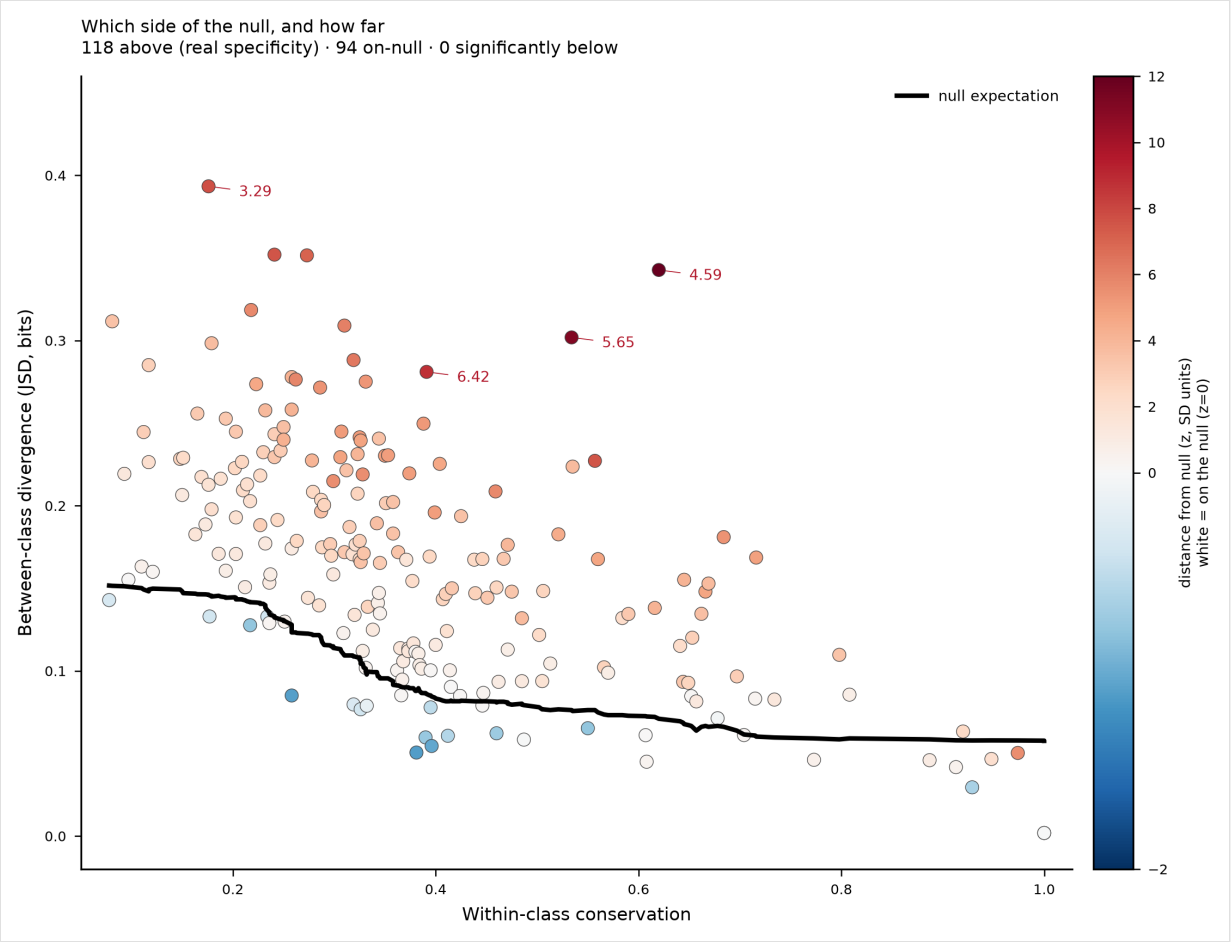

Fig 2. The signature negative trend is mostly the null's shape. Observed conservation–divergence trend ρ = −0.61; permutation-null trend ρ = −0.90. Real signal is the per-position excess above the null (right), not the negative slope.

The null's shape recurs across families

This is not a GPCR peculiarity. Repeating the null construction on four independent families (Fig 3) shows the null trend is strongly negative in every one, from ρ_null = −0.78 (HLA) to −0.97 (opsin), and in three of four the null *explains ≥100%* of the observed trend — meaning the real signal partly *cancels* the null shape rather than creating it.

Family	Grouping (classes)	Observed ρ	Null ρ	Null explains	Frac above null
GPCR	coupling (3)	−0.607	−0.897	148%	0.56
opsin	spectral (5)	−0.911	−0.968	106%	0.72
NR-DBD	response element (2)	−0.794	−0.872	110%	0.78
HLA	supertype (6)	−0.995	−0.780	78%	0.18

Each family's null floor has a different height and roughness, set by its number of classes and its per-class sample size. **This is the mechanistic root of the companion P2's finding that a divergence threshold cannot be transferred between families** — and the two papers are complementary here: P2 shows the floors are *different* (so thresholds don't move);

this paper shows the signature slope itself *comes from* the floor. HLA is the one outlier (null explains 78%, lowest excess fraction at 0.18), most likely because it has the smallest panel (41 sequences across 6 groups); the companion P2 addresses this directly with an IPD-IMGT/HLA scale-up. We report HLA as a caveat, not as support for the universality claim.

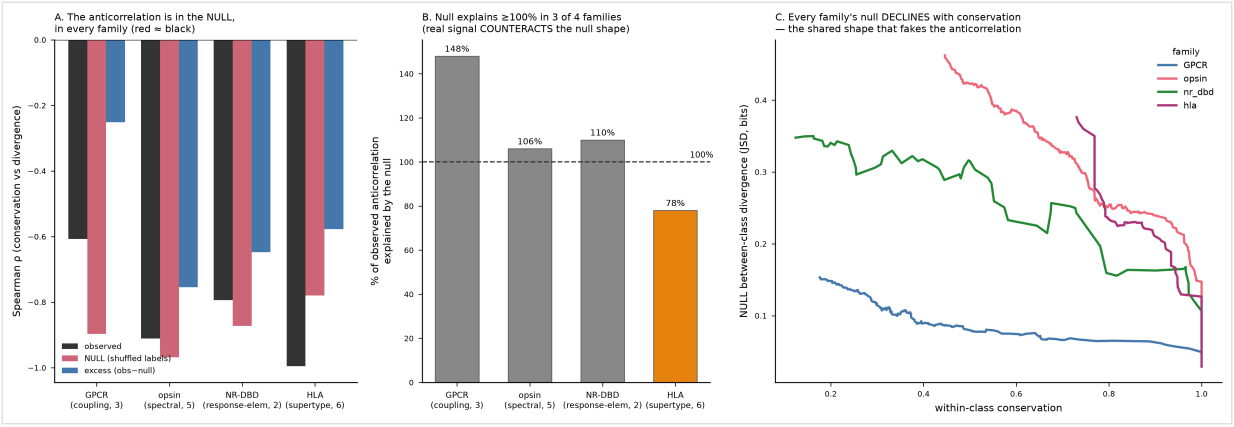

Fig 3. The null's negative shape recurs across families. Four families, all with strongly negative null trends (−0.78 to −0.97); three of four have null_explains ≥ 100%, HLA the exception at 78%. The floor's family-specific height is why divergence thresholds do not transfer (companion P2).

Three-quarters of the PLM's per-position signal is phylogeny

If the intuition detective is to be believed, we must first subtract its homology shortcut. Replacing random 5-fold cross-validation with clade-aware cross-validation — keeping related receptors together so the probe cannot win by recognizing evolutionary cousins — decomposes the PLM's per-position signal cleanly (Fig 4). The random-split probe reaches a median balanced accuracy of 0.66; the clade-aware probe drops to 0.41; chance is 0.33. So of the PLM's total per-position signal above chance (median 0.33), phylogeny accounts for 0.24 (about 74%) and pure coupling for 0.08 (about 24%). This is a direct, quantitative instance of the "PLMs encode homology, not mechanism" critique: at the per-position level, three-quarters of what this PLM appears to know about coupling is evolutionary relatedness. The remaining quarter is real but thin, and we treat it as thin throughout.

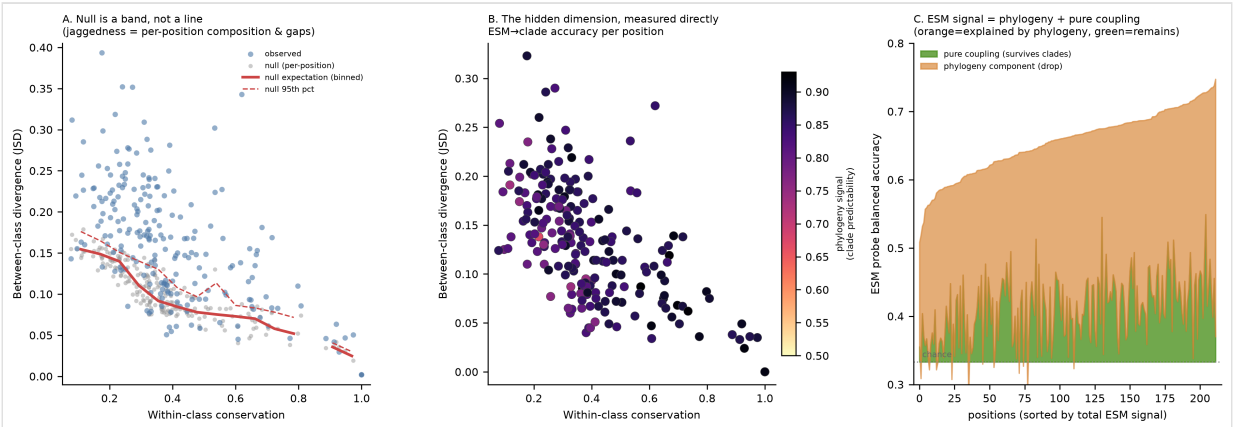

Fig 4. Phylogenetic decomposition of the PLM's per-position signal. Random-split balanced accuracy 0.66 → clade-aware 0.41 (chance 0.33): phylogeny ≈ 74%, pure coupling ≈ 24% of the above-chance signal. The pure-coupling residual is real but thin.

Where they disagree, intuition aligns with the physical interface

The two detectives disagree (ρ = 0.17); the gold standard says the disagreement is directional. Our gold standard is the set of positions that physically contact Gα in the 212 solved GPCR–Gα complex structures (GPCRdb, generic-numbered): 20 interface positions and 7 class-discrimination positions fall within the spectrum. Testing each detective for enrichment on these positions (Fig 5, Mann–Whitney):

Detective	On interface	On class-discrimination
Physical evidence (net divergence)	p = 0.18 (n.s.)	p = 0.73 (n.s.)
Intuition PLM (L30)	p = 0.003	p = 0.042
Intuition PLM (L21)	p = 0.13	p = 0.45

Where the detectives disagree, the PLM (at layer 30) is significantly enriched on the true structural coupling interface and the evidence detective is not. Three honest boundaries travel with this result, and belong beside it, not in a footnote. First, the structural interface is only a *partial* gold standard — it captures direct Gα contacts but misses allosteric and second-shell positions — so the evidence detective's non-enrichment is *not* evidence it is wrong; it may be reading an allosteric layer the interface cannot see. Second, the PLM enrichment is a **distribution-level shift** (Mann–Whitney), not a **top-list win** (Fisher exact on the top-20 is not significant): a stable weak signal, not a ranked-list victory. Third, the interface positions are only *slightly* conservation-biased (14/20, p = 0.03) and their divergence is *not* reduced (9/20, p = 0.83) — so this must not be restated as "the conserved scaffold does not diverge."

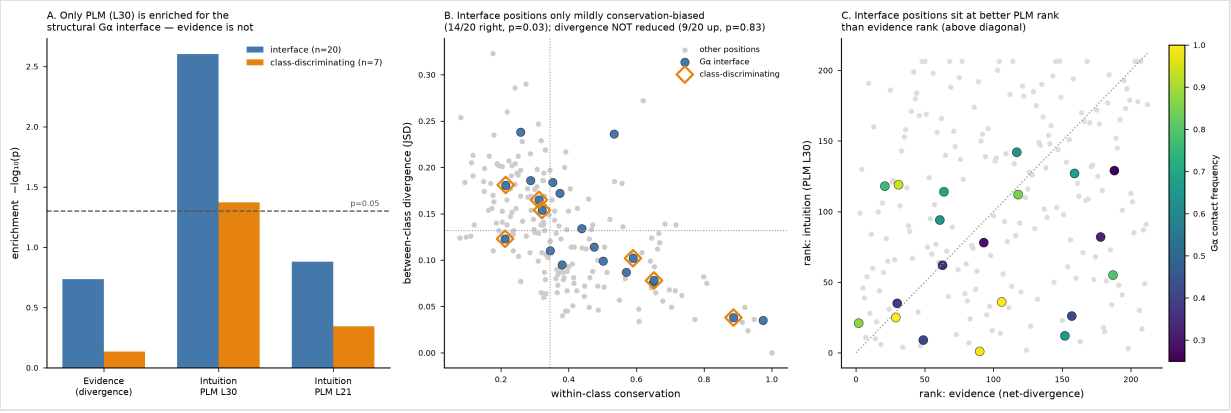

Fig 5. Gold-standard anchoring at the points of disagreement. The intuition detective (PLM L30) is enriched on the physically measured Gα interface (p = 0.003) where the evidence detective is not (p = 0.18) — a distribution-level shift, not a top-list win, against a partial gold standard.

The conclusions are robust — and one is actively falsified

Repeating the decomposition across four ESM layers (6/15/21/30) and two probe types (linear, MLP) tests whether the story is a layer-specific accident (Fig 6). Two claims hold across every layer: phylogeny dominates the per-position signal (68–87% share), and the pure-coupling residual is orthogonal to the divergence axis. But a third claim — that the residual is biased toward conserved positions — **appears only at layer 30** ($p = 0.006$) and is not significant at layers 21 or 15. We report it as a layer-30 artifact, not a general rule. This active falsification narrows an earlier, more exciting reading ("the conserved scaffold carries coupling information") down to what the data actually support: a thin residual, orthogonal to the divergence axis, with no robust conservation bias. The signal plateaus at layer 21 (clade-aware 0.442, above layer 30's 0.416), so the main result uses layer 21; linear and MLP probes are nearly identical (0.416 vs 0.407), showing the coupling information is linearly readable.

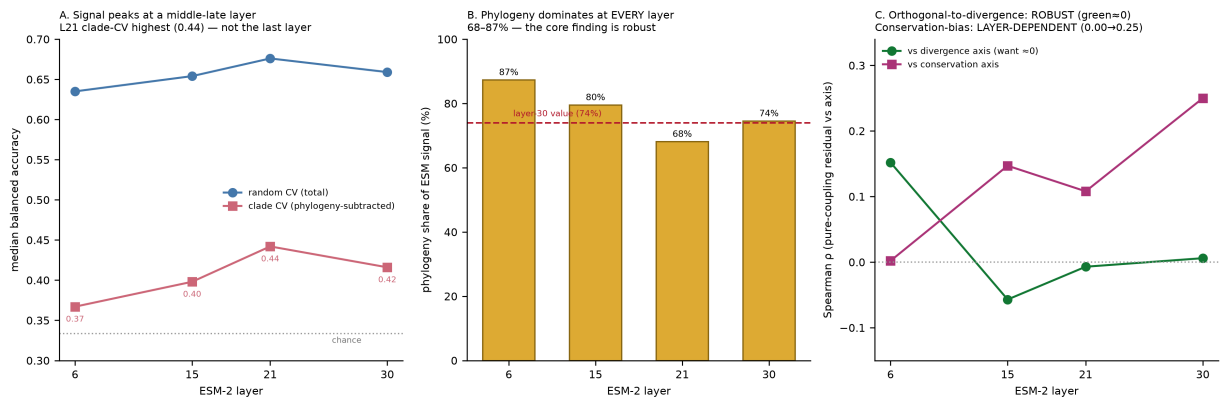

Fig 6. Layer and probe robustness. Phylogeny dominance (68–87%) and residual–divergence orthogonality hold across all layers; the residual's conservation bias appears only at layer 30 ($p = 0.006$; n.s. at L21/L15) and is reported as an artifact. Signal peaks at layer 21.

A negative control: where the biology is unreadable, both detectives fail together

A framework that finds a directional disagreement should also show what happens when there is nothing to find. The companion P2 established opsin spectral tuning as a clean miss — its spectral classes are branch-confined and its tuning is distributed across epistatic sites. We ran both detectives on opsin (144 opsins, 348 positions, five spectral classes; gold standard the 15 known spectral-tuning sites) exactly as on GPCR (Fig 7; PLM side computed with ESM-2 650M, layer 33 — a larger model than the GPCR main analysis, so if anything a generous test of the intuition detective).

Neither detective recovers the tuning code. The PLM puts 4 of 15 tuning sites in its top 15% (0 in the top 5%; top-list Fisher $p = 0.11$, n.s.); the evidence detective puts 4 of 15 in the top 15% (2 in the top 5%; Fisher $p = 0.17$, n.s.). Both show only a weak distribution-level shift (Mann–Whitney $p = 0.003$ and 0.006), neither a top-list win. And the two detectives are essentially uncorrelated here ($\rho = 0.005$). This is the expected signature of a genuine miss: when the biology is unreadable by either method, the two independent detectives fail *together*

and agree with each other no better than chance — reinforcing the companion P2's point that the applicability limit is a property of the family, not of any one detector. One honest note: the PLM's phylogeny-corrected pure-spectral signal on opsin is not zero (clade-aware median 0.41 vs chance 0.20), somewhat higher than the GPCR pure-coupling residual — but it is diffuse and does not land on the textbook tuning sites, which is precisely opsin's known distributed-encoding failure. The miss is at the top-list level, not an absence of all signal.

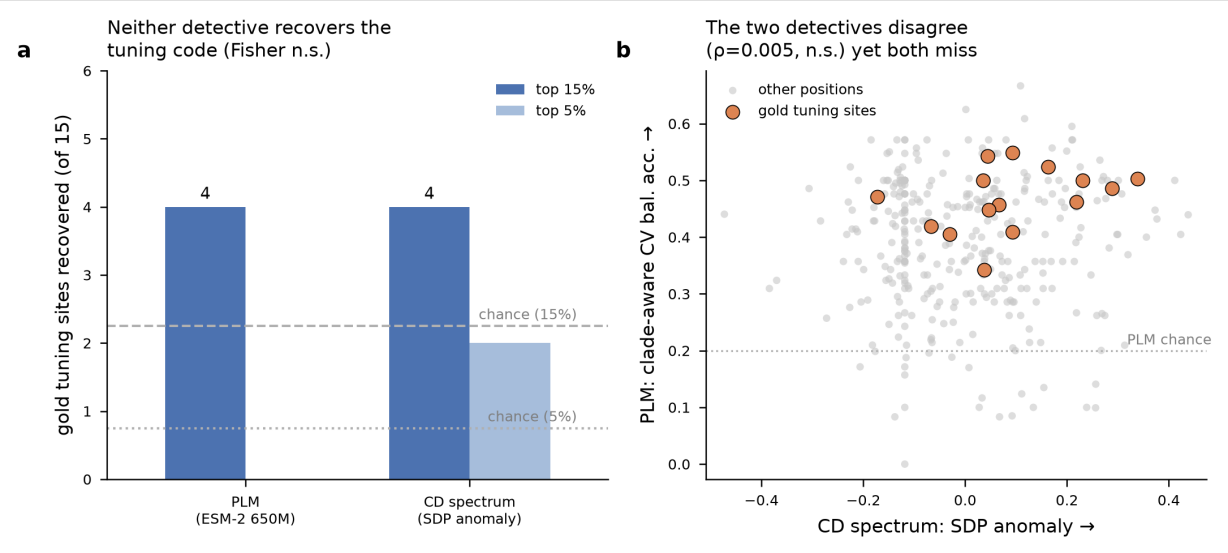

Fig 7. Opsin negative control. On a family the companion P2 established as unreadable, both detectives miss the known spectral-tuning code (top-list Fisher n.s. for each) and disagree with each other (ρ = 0.005) — the expected signature when the biology is genuinely unreadable.

Discussion

Two methods now read functional specificity from protein sequence in fundamentally different ways, and this paper is a protocol for making them testify against each other productively. The value is not a new score — the physical-evidence detective is the companion P1's spectrum, and the intuition detective is an ordinary probe on off-the-shelf embeddings — but a way of cross-examining a black-box model's per-position knowledge with an independent evolutionary witness, under the null-model discipline that makes the cross-examination trustworthy.

Three methodological results generalize beyond GPCRs. First, a signature pattern of the evolutionary method — the conservation–divergence negative trend — is mostly the shape of the permutation null, recurring across four families (with HLA the weakest case, null_explains 78% vs 106–148% elsewhere), and only the excess above the null carries signal. Any per-position specificity method that reports a conservation–divergence relationship without a matched null is at risk of reporting the null's shape as a finding. Second, three-quarters of a PLM's per-position "knowledge" of coupling is phylogenetic relatedness, recoverable only by clade-aware cross-validation; random-split accuracy overstates what the model knows about mechanism by roughly fourfold here. Third, at the points where the two methods disagree, the PLM aligns better with the physically measured interface — but weakly, as a distribution shift rather than a ranked-list win, and against a gold standard that is only partial. The honest

headline is that the intuition detective sees *something* on the interface the evidence detective does not, not that it solves coupling.

The negative control matters as much as the positive results. On opsin, a family independently established as unreadable, both detectives fail at the top-list level and disagree with each other no better than chance. That two independent methods fail together is the signature that the limit is biological, not a defect of one detector — the same conclusion the companion P2 reaches from a different direction.

Boundaries. These are stated where each result is made; we collect the load-bearing ones here. The pure-coupling residual is thin (24% of the PLM's above-chance signal) and was measured with a single PLM size (150M for GPCR; a 650M control on opsin); whether a larger model thickens the residual is untested. The structural interface is a partial gold standard, so the evidence detective's non-enrichment is not a verdict against it. The cross-position comparability of net divergence rests on a Gamma-fit tail probability (validated at $\rho = 0.995$ against the empirical null, goodness-of-fit 4×10^{-4}) but still assumes the JSD null is Gamma-approximable. And the "null shape = f(number of classes, per-class sample size)" relationship rests on only four family data points with the two variables acting in opposite directions, so we state it as a mechanism, not a fitted law.

Relation to the companion preprints. P1 introduces the conservation–divergence spectrum and establishes that it recovers known specificity determinants. P2 asks when it can be trusted and gives a prospective, two-dimensional applicability criterion. This paper takes the instrument to its hardest test — a position-by-position confrontation with a black-box model on the same family — and supplies the null-model and phylogenetic-subtraction discipline that such a confrontation requires. Two threads run through all three: every claim of signal is tested against a permutation null before it is believed, and the null itself is treated as the object of study, not a nuisance. Result §"the null's shape is universal" is the quantitative bridge between this paper and P2's threshold-transfer finding. All three are posted together as companion preprints (Research Square, 2026; cross-referencing DOIs to be added on posting).

What the spectrum does not predict. A distinct boundary concerns what a high S_sdp position does *not* imply, tested on olfactory receptors where within-repertoire variability is abundant. Two orthogonal negatives bound it. First, on human deorphanization data (M2OR; 114 deorphanized receptors, 581 within-family pairs), pocket-residue identity at S_sdp-nominated positions tracks odor tuning only weakly and only in the two largest families, and in leave-one-receptor-out prediction it adds nothing over whole-sequence identity (pocket AUC 0.692 vs sequence 0.693; random-position null 0.691; enrichment $p=0.398$). Second, on human–mouse divergence (1,761 mouse ORs on the human frame), S_sdp correlates with per-position Jensen–Shannon divergence ($\rho=+0.20$) only because it scores high-entropy positions: the association vanishes under partial correlation controlling for conservation or entropy ($\rho=-0.06$, $p\approx0.42$) and under entropy-stratified matching (Δ median JSD ≈0). This is the same null-discipline lesson as the conservation–divergence trend itself, turned on the score's own output: because S_sdp is constructed to favour low-conservation positions, it will spuriously track *any* external variable that also favours them unless that

confound is subtracted. The engine nominates subfamily-discriminating positions — its designed purpose — but those positions do not individually predict quantitative functional tuning, nor carry cross-species plasticity information beyond conservation.

Methods

Data and alignment. 219 class A GPCRs with primary coupling assignments (GtoPdb / GPCRdb), aligned to the GPCRdb generic-numbering frame; 212 positions present across the panel. Coupling classes Gs, Gi-0, Gq-11. Opsin control: 144 opsins on the bovine-rhodopsin coordinate frame, 348 positions, five spectral classes (RH1, RH2, SWS1, SWS2, LWS), 15 gold-standard spectral-tuning sites.

Physical-evidence scores. S_sdp is the companion P1's headline anomaly_score: between-class Jensen–Shannon divergence minus a global linear fit of divergence on within-class conservation across the 212-position family, i.e. the residual above the family-wide conservation–divergence trend line (P1's add_anomaly(); see P1 Methods). This is P1's primary specificity-anomaly score, not its auxiliary sdp_score (a conservation × divergence product P1 also reports internally); the two correlate at only $\rho = 0.78$ on the same positions, so this substitution changes S_sdp's numeric values from an earlier internal draft that had carried sdp_score forward under the S_sdp label — the $\rho = 0.88$ (S_sdp vs S_net) and $\rho = 0.17$ (S_net vs S_plm, unaffected) reported throughout this paper reflect the corrected anomaly_score-based S_sdp. S_net subtracts the permutation-null floor and expresses the excess as a Gamma-fit tail probability ($-\log_{10} p$), validated against the empirical null at $\rho = 0.995$.

Permutation null. Two independent permutation-null pipelines are used in this paper and are not interchangeable. (1) The per-position GPCR floor, excess, and z-score behind Fig 1/Fig 2 and the "side of the null" classification (null_sides.csv, built from null_curve.csv) come from shuffling coupling labels while preserving amino-acid composition; the exact permutation count used to build null_curve.csv is not separately re-verified here. (2) The Gamma-fit tail probability behind S_net (comparable_scores.csv) is built independently, shuffling coupling labels 2000× per position with amino-acid composition preserved, and its own empirical-vs-Gamma agreement is validated directly ($\rho = 0.995$ against the empirical null). The conservation–divergence null trend reported in Results ($\rho_{null} = -0.90$) is computed from pipeline (1). null_explains = ρ_{null} / ρ_{obs}; frac_above_null = fraction of positions with $z > 2$. The four-family cross-family comparison (GPCR, opsin, NR-DBD, HLA; Fig 3) uses a separate implementation of the same permutation-null logic with 200 shuffles per family; we report its ρ_{null} values as computed, without a separate convergence check at this budget.

Intuition (PLM) probe. ESM-2 per-position embeddings (esm2_t30_150M_UR50D, 640-dim; layer 21 for the GPCR main result, layers 6/15/21/30 for the robustness scan). For each position, a logistic-regression probe (standardized features, $C = 0.5$, class-balanced) predicts coupling class from the embeddings of the receptors present at that position, scored by balanced accuracy. Random 5-fold CV gives the total per-position signal; clade-aware CV (leave-one-clade-out, clades from sequence identity) gives the pure-coupling signal after phylogenetic subtraction. Pure signal = (clade-aware balanced accuracy – chance), clipped at zero. The opsin control uses ESM-2 650M (esm2_t33_650M_UR50D, layer 33) with leave-

one-clade-out CV on sequence-identity clades. The main-result numbers (Fig 4: random-split 0.66, clade-aware 0.41) come from the original layer-21 pipeline; the independent four-layer robustness re-scan (Fig 6, layer_scan_summary.csv) recovers the same layer at slightly different values (random-split 0.676, clade-aware 0.442) from a separately-run probing pipeline — we have not traced the exact source of this small numeric gap (e.g., fold assignment, probe refit, or minor preprocessing differences between the two runs). The two runs agree on every qualitative claim, but Fig 4's and Fig 6's own numbers should each be read from their own table, not interchanged.

Gold-standard anchoring. Interface positions are receptor residues contacting Gα in 212 solved GPCR–Gα complexes (GPCRdb); class-discrimination positions are the subset that differ by coupling class. Enrichment is tested by Mann–Whitney (distribution-level) and Fisher exact on the top-20 (top-list level). Opsin gold standard: 15 published spectral-tuning sites; recall counted in the top 15% / top 5% of each detector's score, top-list enrichment by Fisher exact.

Software and data. ESM-2 [2] via fair-esm; probes via scikit-learn [5]. All master tables, per-position scores, null distributions, and figure-generation code are provided as supplementary datasets (see companion data manifest).

Declarations

Competing interests. The author declares no competing interests.

Funding. This work received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors; H.S. is an independent researcher with no institutional affiliation.

Author contributions. H.S. designed and performed all analyses and wrote the manuscript.

AI assistance. An AI assistant (Claude Science, Anthropic) was used to assist with data analysis, figure generation, and manuscript drafting under the author's direction and review; the author takes full responsibility for the content.

Data availability. All data underlying the findings are within the manuscript and its supplementary datasets. Analysis code (per-position probes, permutation null, phylogenetic decomposition, gold-standard anchoring) is archived on Zenodo under the concept DOI https://doi.org/10.5281/zenodo.21449958 (always resolves to the current version; shared with companion preprints P1 and P2, which use the same conservation–divergence spectrum instrument).

References

All DOIs below were confirmed via Crossref API lookup (api.crossref.org/works/{doi}) to resolve to a metadata record whose title and author list match the citation as printed.

1. Matic M, Singh G, Carli F, De Oliveira Rosa N, Miglionico P, Magni L, Gutkind JS, Russell RB, Inoue A, Raimondi F. PRECOGx: exploring GPCR signaling mechanisms with deep protein representations. Nucleic Acids Res. 2022;50(W1):W598–W610. doi:10.1093/nar/gkac426

2. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379(6637):1123–1130. doi:10.1126/science.ade2574

3. Tule S, Foley G, Bodén M. Do protein language models learn phylogeny? Brief Bioinform. 2024;26(1):bbaf047. doi:10.1093/bib/bbaf047

4. Lupo U, Sgarbossa D, Bitbol AF. Protein language models trained on multiple sequence alignments learn phylogenetic relationships. Nat Commun. 2022;13:6298. doi:10.1038/s41467-022-34032-y

5. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–2830.

Data and code

All per-position score tables (three_detectives.csv, decomposition.csv, null_sides.csv, comparable_scores.csv), cross-family null summary (crossfamily_null_summary.csv), gold-standard anchoring (anchor_enrichment.csv, interface_goldstandard.csv), layer/probe robustness (layer_scan_summary.csv), and the opsin negative control (opsin_plm_probe.csv, opsin_negative_control_summary.csv) are provided as supplementary datasets. Companion preprints P1 and P2 (Research Square, 2026; DOIs to be added on posting).