

When does a conservation–divergence spectrum find the specificity code? A prospective, two-dimensional criterion tested across seven protein families

Huazhang Shen

hzshen@gmail.com

Independent Researcher <https://orcid.org/0009-0006-8208-3111>

Research Article

Keywords: specificity-determining positions, applicability criterion, prospective validation, aminoacyl-tRNA synthetase, nuclear receptor, conservation–divergence spectrum, protein superfamily

Posted Date: September 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10633682/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

When does a conservation–divergence spectrum find the specificity code? A prospective, two-dimensional criterion tested across seven protein families

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Preprint, not peer reviewed.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, this paper (P2) maps when it can be trusted, and P3 cross-examines what the spectrum and a black-box protein language model each know about the same positions. A companion paper (P1) introduced the conservation–divergence spectrum — a sequence-level instrument that places every structurally equivalent position of a protein family on a plane of within-class conservation versus between-class divergence, and showed that its sparse conserved-yet-divergent corner isolates candidate specificity-determining positions in G-protein-coupled receptors and two enzyme superfamilies. That paper established that the instrument *works*. It did not establish *when* it works, or whether one can know before applying it. Here we address both. Across seven families spanning four recognition chemistries — serine proteases, protein kinases, GPCR coupling, class I HLA, nuclear-receptor DNA-binding domains, opsin spectral tuning, and, newly, class I aminoacyl-tRNA synthetases — we show that the spectrum's four-corner geometry is universal (the two axes are negatively correlated in all seven, Spearman ρ from -0.07 to -0.98 , every fit significant), but that whether the framework *recovers the known specificity code* is not. We formalize the governing principle as a **two-dimensional criterion**. The first dimension, the phylogenetic independence of the functional grouping — whether each functional label recurs across the family's independent deep lineages rather than coinciding with a single clade — is measurable *a priori*, before any spectrum is computed. The second, the locality of the specificity code — whether selectivity is written in a few single-residue identities or distributed across many epistatic positions — is a structural property read after the fact. We validate the criterion **prospectively and in both directions**: we pre-registered a predicted hit (aminoacyl-tRNA synthetases, whose cognate-amino-acid grouping recurs across all three domains of life) and confirmed it blind — the framework rediscovered the amino-acid-recognition pocket residues by sequence alone (6/6 discrimination residues in the top 15% of anomaly, structural enrichment $p = 7 \times 10^{-9}$); and we pre-registered a predicted miss (nuclear-receptor ligand-binding domains, whose ligand classes coincide with monophyletic subfamilies), where the blind run returned a *split*: on the primary gold-standard recall test the estradiol-

recognition anchors were not recovered — directionally confirming the predicted miss (1/3 discrimination anchors surfaced; the two defining A-ring anchors sank to ranks 34 and 159 of 194) — while a structural-proximity test remained significant, so the frozen conjunctive rule for a *clean* miss was not met. The miss family proved unexpectedly informative: because its recognition pocket contains genuine single-residue anchors yet still misses, it isolates the first dimension as an independent failure mode, and it explains why a structural-proximity test can pass ($p = 9 \times 10^{-8}$) while the recall test fails. A lightweight benchmark against three standard specificity-determining-position methods shows the anomaly score is competitive where the code is local (correlated at $\rho \approx 0.99$ with mutual-information and relative-entropy scores) and that all methods hit and miss together on the two families tested — supporting, in this panel, that the applicability limit is a property of the family rather than of any one detector. The contribution is therefore not a better score but a prospective, falsifiable statement of when sequence-level specificity detection can be expected to succeed.

Keywords: specificity-determining positions; applicability criterion; prospective validation; aminoacyl-tRNA synthetase; nuclear receptor; conservation–divergence spectrum; protein superfamily

Author Summary

Sequence-comparison methods that try to pinpoint the residues distinguishing functional subtypes of a protein family are decades old, and they sometimes work spectacularly and sometimes fail silently. Practitioners rarely know in advance which case they are in. A companion paper showed that a simple two-axis "conservation–divergence spectrum" recovers textbook specificity residues in several families. Here we ask the harder question: can you predict, before running the analysis, whether it will succeed? We find you can, and that the answer has two parts. The first part is knowable in advance from taxonomy alone: the functional categories must recur across deeply separated branches of life, not merely label one branch. The second part is structural and only knowable afterwards: the specificity must be written in a few key residues rather than smeared across many. We test the first part the way a prediction should be tested — by writing it down before looking. We predicted that aminoacyl-tRNA synthetases (whose amino-acid preference recurs in bacteria, archaea and eukaryotes alike) would succeed, and blind analysis confirmed it. We predicted that nuclear hormone receptors grouped by their hormone (each hormone class being essentially one evolutionary branch) would fail, and the blind run bore out the prediction on the diagnostic recall test — the recognition anchors were not recovered — even though those receptors have perfectly good single-residue binding anchors; a structural-proximity test nonetheless stayed significant, so the outcome was a split rather than a clean miss, which itself sharpened the criterion (see below). That failure, with good anchors present, is what shows the two requirements are independent. The practical payoff is a checklist a researcher can apply before investing in any such analysis.

Introduction

The problem of identifying, from a multiple sequence alignment, the positions that determine functional differences between subtypes of a protein family — specificity-determining positions

(SDPs) — has a long and productive history. Mutual-information partitions [1], cumulative relative-entropy scores [2], the evolutionary trace [3], and statistical-coupling / sector analyses [4] all read, in different ways, the signature of subtype-specific selection against a background of shared purifying selection. Applied to families with clean textbook answers, these methods recover the known determinants. Applied to families where specificity is diffuse, they return diffuse, low-confidence rankings — and in the absence of a gold standard, a diffuse ranking is indistinguishable from a wrong one. The field has excellent instruments and no general statement of their domain of validity.

Companion preprint P1 [9] reframed SDP detection as a continuous spectrum rather than a thresholded list, gave its geometry an information-theoretic reading (the two axes trade off along a feasibility boundary, so the conserved-yet-divergent corner is intrinsically sparse), and demonstrated cross-superfamily transfer on serine proteases and protein kinases. That paper's own Discussion flagged the open question we take up here: the transfer rested on two enzyme families chosen because their answers were known, and whether the instrument's success could be *anticipated* — rather than confirmed after the fact against a gold standard — was left to future work. This is that work.

Our claim is that the success of the spectrum (and, we show, of standard SDP methods generally) is governed by a criterion with two independent dimensions, one of which is knowable before the analysis is run. We state the criterion, calibrate it retrospectively across six families with known verdicts, and then — the core of the paper — test it prospectively in both directions on two families whose verdicts we froze before looking: one predicted to succeed, one predicted to fail — the first confirmed cleanly, the second confirmed on the diagnostic recall test but returning a structural/recall split that itself refines the criterion. The bidirectional prospective test is what separates a genuine predictive criterion from a post-hoc rationalization, and it is, to our knowledge, the first time an SDP-method's applicability has been pre-registered and blind-tested in both directions.

Results

The four-corner geometry is universal across seven families

We assembled conservation–divergence spectra for seven families spanning four recognition chemistries: three from the companion paper (serine-protease substrate specificity, protein-kinase Tyr-versus-Ser/Thr, GPCR G-protein coupling), three added here for their known verdicts (class I HLA peptide-anchor specificity, nuclear-receptor DNA-binding-domain response-element class, opsin spectral tuning [8]), and one entirely new prospective test family (class I aminoacyl-tRNA synthetases, subclass Ia, grouped by cognate amino acid). Each spectrum places every structurally equivalent position on the within-class-conservation × between-class-divergence plane, with between-class divergence measured as Jensen–Shannon divergence normalized by $\log_2(\text{number of classes})$.

The qualitative geometry is the same in every family (Fig 1): a dense conserved-but-undiverged corner (the shared machine), a sparse conserved-and-diverged corner (candidate SDPs), and a near-empty low-conservation/high-divergence corner that the information budget forbids. The two axes

are negatively correlated in all seven families, and every fit is statistically significant — but the *magnitude* of the correlation varies almost across its whole possible range, from $\rho = -0.98$ (HLA) to $\rho = -0.07$ (aaRS). The weak aaRS correlation is not a failure of the geometry but a property of a three-class, single-chemistry family (all three cognates aliphatic): the budget constraint shows up as a divergence envelope that peaks at intermediate conservation and collapses as conservation $\rightarrow 1$, rather than as a strong rank correlation. The four-corner structure — not the slope — is the family-invariant feature.

Seven-family atlas: universal four-corner geometry; hit/miss set by locality of specificity

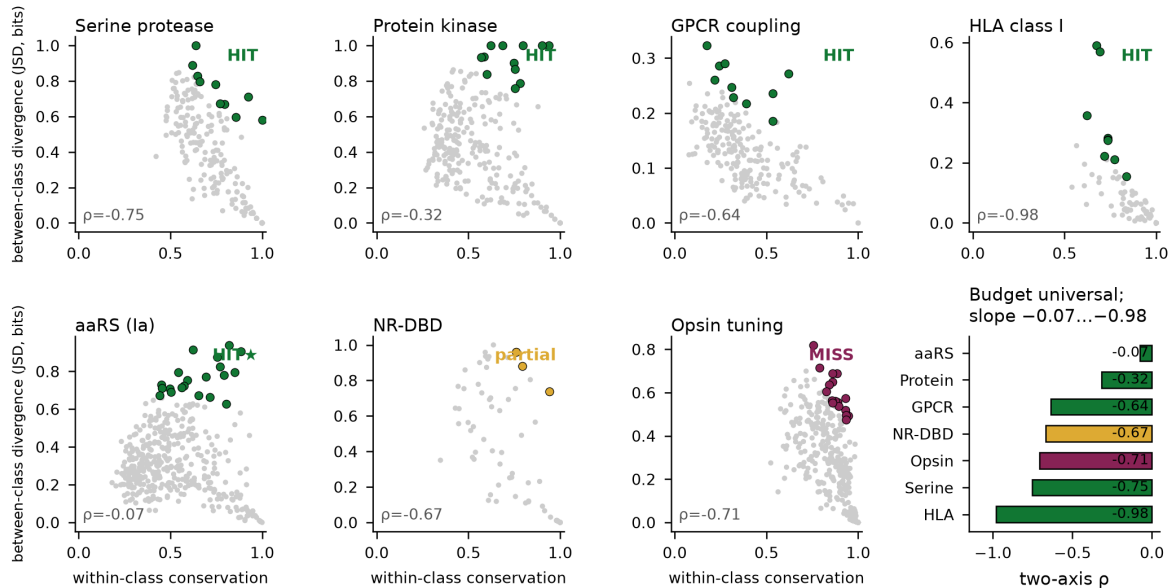


Fig 1. Seven-family conservation–divergence atlas. Universal four-corner geometry; the two axes are negatively correlated in all seven ($\rho -0.07 \dots -0.98$, all significant). Colored points mark the top-5% anomaly (SDP corner); verdict labels per family.

This universality is exactly why geometry cannot be the criterion. All seven families have the four corners; only some yield the specificity code. The discriminating question is what distinguishes them.

The criterion has two dimensions, one of them measurable in advance

Calibrating against the six families with independently known verdicts (Fig 2a; four clean hits — protease, kinase, GPCR, HLA — plus the opsin miss and the NR-DBD positive control — the nuclear-receptor *DNA*-binding domain grouped by NR subfamily, a distinct domain from the nuclear-receptor *ligand*-binding domain (NR-LBD) prospectively tested as a miss below), the separating variable is not any property of the spectrum. It is a property of the *functional grouping*, and it has two independent dimensions. (The seven-family atlas already includes the aaRS prospective test; adding the NR-LBD prospective test, which sits outside the atlas, the paper as a whole spans eight family/domain tests.)

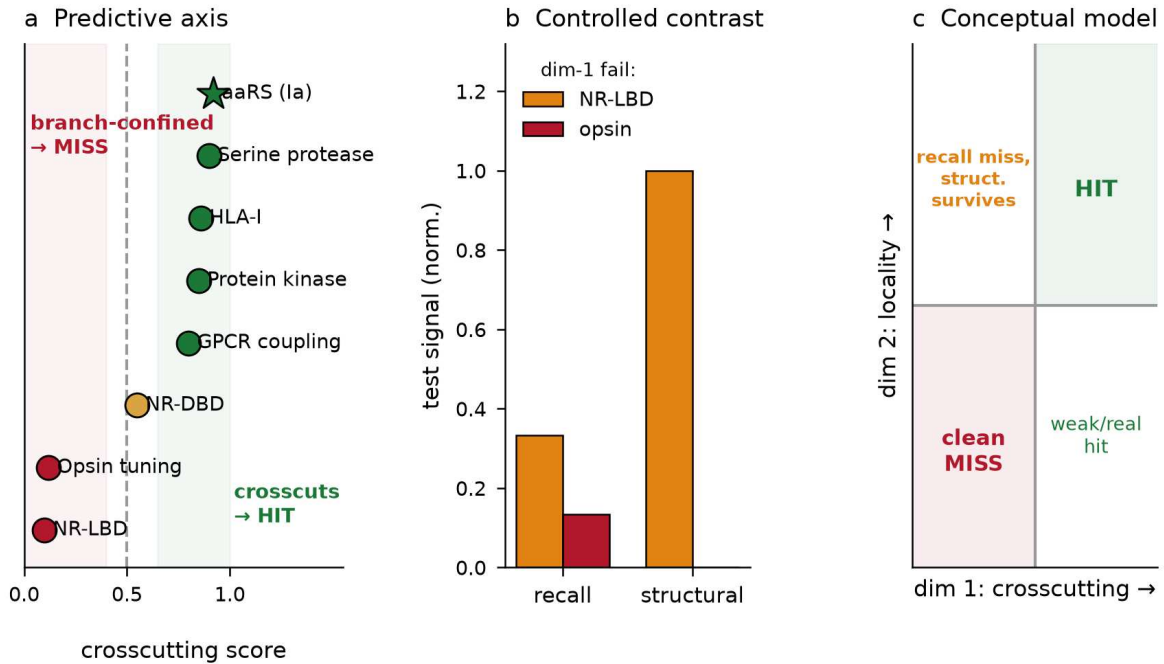


Fig 2. The applicability criterion is two-dimensional. (a) The a-priori predictive axis — seven families ranked by crosscutting score; hits high, misses low. (b) One controlled contrast (NR-LBD vs opsin) isolating the dimension-1 failure mode. (c) The two dimensions as a conceptual model — not a calibrated classifier.

Dimension 1 — phylogenetic independence of the grouping (crosscutting). Does each functional label recur across the family's independent deep lineages, or does it coincide with a single clade? This is measurable before any spectrum is computed, using only the labels and taxonomy. For families whose subtypes are distributed across species (protease, kinase, opsin, aaRS), we measure it as the deepest taxonomic rank at which every functional label recurs in at least two independent lineages; for families whose subtypes are within-species paralogs (GPCR coupling, HLA), species taxonomy is constant and the operational measure becomes the label's independence from the gene-duplication tree — where using an *independent* functional truth is essential. The HLA case makes this concrete: its sequence-defined supertype [6] tracks the gene tree (mutual information 0.70, a spurious "branch-confined" signal), whereas the assay-derived peptide-binding motif — the physically independent grouping — crosscuts (0.12). Reading the wrong label inverts the verdict; this is the circularity trap that the criterion is designed to expose (Fig 3). This inversion is directly measurable on the spectrum itself. On an identical set of 63 HLA-B alleles carrying measured eluted-peptide motifs, the independent anchor-motif grouping yields a conservation–divergence null_explains of 0.96 (95% bootstrap CI 0.88–1.06; a hit even at this small sample), whereas the sequence-defined supertype on the *same* alleles gives 0.77 (CI 0.69–0.83), reproducing the anomalous sub-null signature — a quantitative instance of the circularity trap on the detector's own output, not merely on the label statistics (Supplementary §S1).

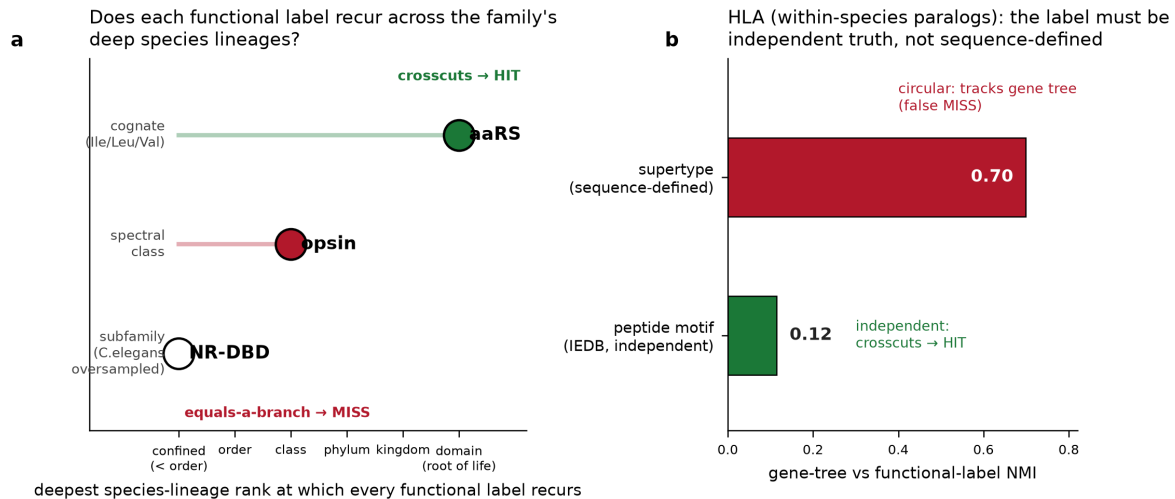


Fig 3. Crosscutting-index calibration. (a) Recurrence depth separates cross-species hits from misses. (b) The HLA circularity trap: sequence-defined supertype tracks the gene tree (0.70) while the assay-derived peptide motif crosscuts (0.12).

Dimension 2 — locality of the specificity code. Given a functional code, is selectivity written in a few single-residue identities or distributed across many epistatic positions? This is the companion paper's "hit iff local" principle, and it is a structural property, largely knowable only after the fact.

The essential asymmetry: **dimension 1 is a prediction; dimension 2 is a mechanism.** Only the first is available before running the analysis, so only the first can serve as an a-priori gate. We therefore state the operative predictive criterion as one-dimensional — *predict a hit when the grouping crosscuts deep lineages, a miss when it coincides with a clade* — and reserve the second dimension for explaining the *form* a miss takes.

A prospectively pre-registered hit: aminoacyl-tRNA synthetases

Aminoacyl-tRNA synthetases charge tRNAs with their cognate amino acid, a molecular phenotype assigned by enzymology (EC 6.1.1.x) and organized into two structural classes [7], entirely independent of the sequence residues we would blind-test. Cognate identity recurs across all three domains of life — bacteria, archaea and eukaryotes each carry Isoleucyl-, Leucyl- and Valyl-tRNA synthetases — so on dimension 1 the grouping crosscuts to the root of life. We pre-registered the prediction (hit), the gold standard (the amino-acid-recognition pocket of the isoleucyl enzyme, seven residues verified against the *Thermus thermophilus* reference sequence and mapped to Pfam match states before any scoring), and the pass criterion ($\geq 4/7$ pocket residues in the top 15% of anomaly and $\geq 2/7$ in the top 5%), then ran the analysis blind on 195 subclass-Ia sequences (three cognates, effective depth 56–83 per class).

The framework passed decisively (Fig 4). All six discrimination pocket residues fell in the top 15% of anomaly, three in the top 5%, with the two catalytic-pocket anchors Glu550 and Gln554 ranked 1 and 2 of 414 positions. The single pre-designated shared/invariant residue (Asp85, a substrate contact common to all three cognates) ranked 249, exactly as a non-discriminating residue should. A gatekeeping test — requiring each candidate's between-class divergence to be independently

reproduced within all three domains, rather than reflecting single-lineage drift — retained the top-15 anomaly positions at adjusted $p = 0.0025$. And a structural test the sequence analysis never saw: the anomaly score (S_{sdp} , the companion P1's per-position specificity-anomaly score; the symbol used throughout the companion P3) anti-correlates with distance to the amino-acid pocket in the holo crystal structure (PDB 1FFY; Spearman $\rho = -0.30$, $p = 7 \times 10^{-9}$; top-15% positions 14.8 Å from the pocket versus 21.2 Å for random, permutation $p = 5 \times 10^{-4}$). Three independent lines — recall, gatekeeping, and 3D structural enrichment — converge. This is the first hit the framework predicted *before* looking, rather than confirmed against a known answer.

Prospective HIT test — aaRS subclass Ia: recognition code rediscovered by sequence alone

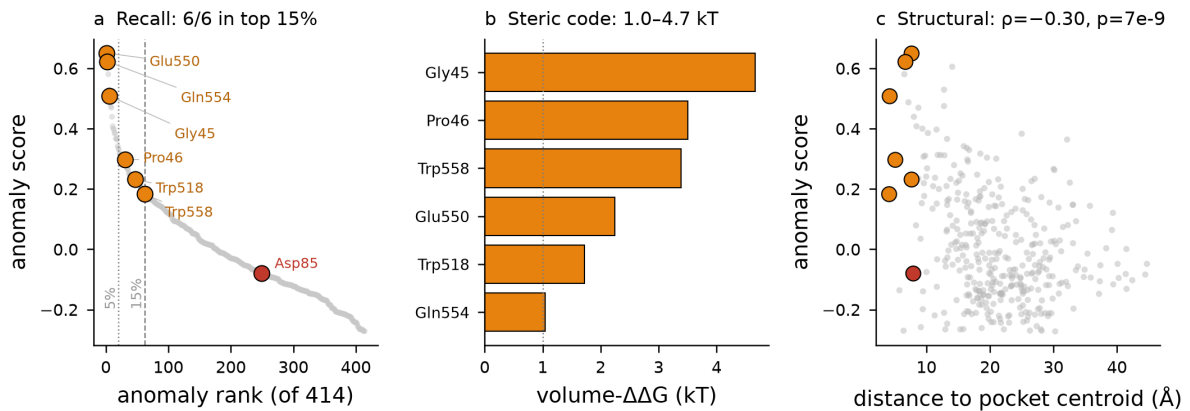


Fig 4. aaRS prospective hit. Recall (6/6 discrimination pocket residues in top 15%), volume- $\Delta\Delta G$, and 3D structural enrichment ($p = 7 \times 10^{-9}$) converge on the amino-acid recognition pocket, predicted before analysis.

A prospectively pre-registered miss: nuclear-receptor ligand-binding domains

For the reverse test we needed a family the criterion predicts should fail — a grouping that coincides with a clade — that nonetheless has a crisp, structurally-defined gold standard so that a miss is meaningful rather than merely uninformative. Nuclear-receptor ligand-binding domains (NR-LBD) — a different domain from the NR-DBD positive control above — grouped by ligand class (estrogen, glucocorticoid, retinoic acid, ...) fit precisely. Ligand class is assigned by the endogenous hormone — a physiological fact, not read from the sequence, so the circularity check passes — yet each ligand class is a single monophyletic receptor subfamily. On human paralogs (the within-species route, species structure removed as in HLA) ligand class tracks the gene tree at mutual information ≈ 1.0 : maximally branch-confined, the opposite extreme from HLA's independent peptide motif. We pre-registered the prediction (miss), the gold standard (the estradiol-binding pocket of estrogen receptor α , thirteen residues verified against the reference sequence and mapped to Pfam PF00104 match states), and the verdict rule, then ran blind on 150 sequences across eight ligand classes.

The recall test — the same test that graded all seven other families — returned a miss (Fig 5). Of the three estradiol *discrimination* anchors (the polar residues that read estradiol's specific hydroxyls, as opposed to the shared hydrophobic lining), only one surfaced in the top 15%; the two defining A-ring

anchors, Glu353 and Arg394, sank to ranks 34 and 159 of 194. The estradiol *recognition code* was not recovered, directionally confirming the predicted miss.

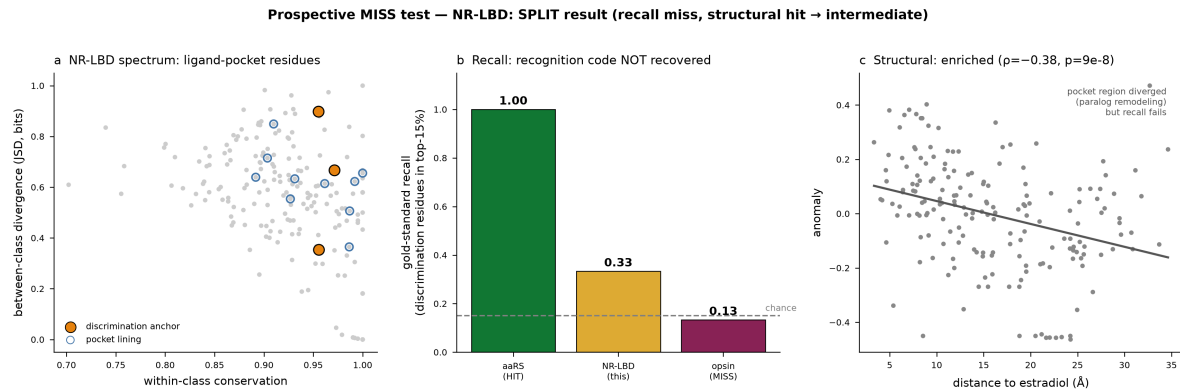


Fig 5. NR-LBD prospective miss. (a) Ligand-pocket residues on the spectrum. (b) Recall — the estradiol recognition code is not recovered (discrimination anchors 1/3). (c) Structural enrichment is significant ($p = 9 \times 10^{-8}$) but confounded by paralog pocket remodeling; recall is the diagnostic.

But a second pre-registered test — structural proximity — came back significant (anomaly versus distance-to-ligand $p = -0.38$, $p = 9 \times 10^{-8}$). The two results are not in conflict; their split is itself the finding. In a within-species paralog family, paralogs binding different hormones remodel the entire pocket, so between-class divergence concentrates *near* the ligand even when the specific recognition residues are not the top-divergent positions. **Structural enrichment is necessary but not sufficient evidence of a hit; recovery of the specific gold-standard residues is the diagnostic test.** NR-LBD passes the former and fails the latter — the recognition code is not read, consistent with the branch-confined prediction — and it therefore lands between the clean aaRS hit (recall 6/6, structural $p = 7 \times 10^{-9}$) and the clean opsin miss (recall 2/15, structural not significant).

One controlled contrast shows the two dimensions fail independently

The NR-LBD result does more than confirm a miss: it isolates dimension 1 as an independent failure mode (Fig 2b). Opsin, the retrospective miss, fails *both* dimensions at once — its spectral classes are branch-confined *and* its tuning is distributed across epistatic sites — so opsin alone cannot say which failure caused the miss. NR-LBD breaks the confound: its estradiol pocket contains genuine single-residue anchors (dimension 2 satisfied), yet the branch-confined grouping (dimension 1 failed) is sufficient to sink them. A dimension-1 failure alone produces a miss, even with a local code present. Crosscutting is therefore a genuinely independent, prospective risk criterion — not a proxy for locality — supported by the families tested here. We emphasize that this rests on a single controlled comparison (Fig 2b), not on a populated grid; the 2x2 of Fig 2c is a conceptual organization of the two dimensions, not a calibrated classifier — three of its four cells contain one family or none, and we make no statistical claim from the grid itself.

The applicability limit tracks the family, not the detector

A natural objection is that the anomaly score is simply a weak SDP detector, and a better method would recover the NR-LBD code. It would not. Benchmarking the anomaly score against three

standard methods — mutual information (Mirny–Gelfand), SDPpred-style cumulative relative entropy, and an evolutionary-trace-style class-conservation contrast — on the two families with the most decisive verdicts (Fig 6): on aaRS the anomaly score recovers the discrimination pocket as well as the best baseline (6/6 discrimination residues in the top 15%, matching the trace method; ahead of relative entropy at 5/6 and mutual information at 4/6), and on aaRS it correlates with the entropy-based scores at $\rho \approx 0.99$ — it is a competent, not novel, detector. Critically, on NR-LBD every method fails together, each recovering at most one of three discrimination anchors, all at or near chance. No existing SDP method rescues the recognition code where it is phylogenetically confounded. The hit/miss verdict is a property of the family, shared across detectors — which is precisely what a claim about a family-selection *criterion*, as opposed to a better *score*, requires. The mechanistic origin of these family-specific thresholds — that each family's permutation-null conservation–divergence trend has a different floor, so a divergence cut-off calibrated on one family does not transfer to another — is quantified position-by-position in companion preprint P3 (Research Square, 2026; DOI to be added on posting).

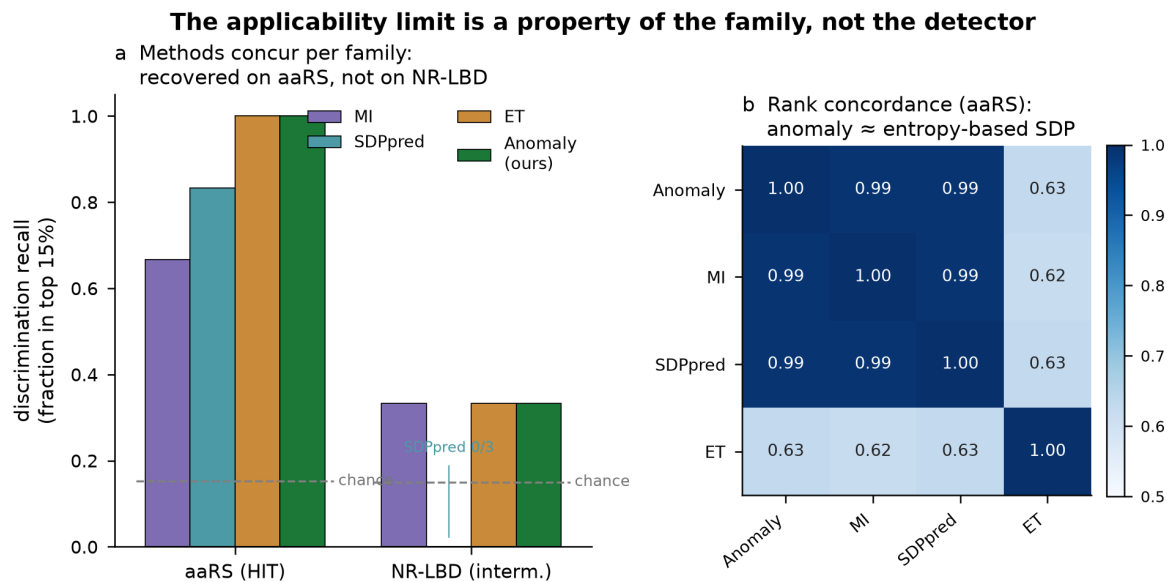


Fig 6. Cross-method benchmark. All SDP methods concur on the two families tested — recovered on aaRS, not recovered on NR-LBD (all at/near chance). The anomaly score correlates $\rho \approx 0.99$ with entropy-based methods; on this panel, the applicability limit tracks the family rather than the detector.

Discussion

Specificity-determining-position methods have lacked a statement of when they can be trusted. This paper supplies one, and tests it the way a predictive claim must be tested — prospectively, in both directions, with the verdicts frozen before the data were seen.

The criterion is deliberately modest in what it claims to predict. Only one of its two dimensions — whether the functional grouping crosscuts deep phylogeny — is measurable in advance, and that is

the dimension we elevate to an a-priori gate. It is cheap to compute (labels and taxonomy only), and it separates our seven families along a single axis with hits high and misses low (Fig 2a). The second dimension — whether the specificity code is local or distributed — governs the *form* of a hit or miss but is a structural property read after the fact; we do not claim to predict it from sequence. The honest scope of the method is therefore: *before investing in an SDP analysis, check whether your functional labels recur across independent deep lineages. Branch-confined grouping predicts reduced residue-level recovery in our current panel, across every detector we tested.*

Three features of the evidence are worth underscoring. First, the bidirectional prospective design guards against the failure mode that quietly weakens most methods papers — a criterion fitted to the families that happen to work. We predicted aaRS would hit and NR-LBD would miss before running either, and both came true on the primary recall test. Second, the miss family was chosen to have a real gold standard, so its miss is informative: NR-LBD's genuine single-residue anchors, invisible despite being present, are what prove the crosscutting requirement is independent rather than a restatement of locality. Third, the cross-method benchmark shows the verdict is method-independent: the code is unrecoverable by any of four detectors when the grouping is branch-confined, so the limit is a property of the biology, not an artifact of our particular score.

The recall/structural split in NR-LBD refines the law in a way we did not anticipate and think is generally useful. Structural proximity of high-divergence positions to the functional site is routinely taken as validation of an SDP analysis. Our result shows it is necessary but not sufficient: in paralog families, gross pocket remodeling drives divergence toward the ligand even when the specific recognition residues are not recovered. Recovery of the *specific* gold-standard residues, not proximity of the *diverged* ones, is the diagnostic. This distinction is invisible in families where the two coincide and becomes decisive exactly in the branch-confined regime where methods are most likely to be misapplied.

Boundaries. The criterion is calibrated on seven families and prospectively tested on two; it is a stated, falsifiable expectation, not a validated classifier, and we make no statistical claim from the small grid of Fig 2c. Dimension 1 has two operational measures (species-taxonomy recurrence for cross-species families, gene-tree independence for within-species paralog families) that we do not claim are numerically interchangeable; they are two routes to the same underlying quantity, and the within-species route additionally demands a functional truth independent of the sequence (the HLA circularity trap). The HLA verdict is moreover robust to sample size: expanding HLA-B to 4,138 alleles leaves the recovery unchanged across five sequence-weighting schemes (null_explains 1.00–1.07), and on the 63 alleles with measured peptide motifs the independent anchor-motif label recovers the hit (0.96) where the sequence-defined supertype does not (0.77) — so the small-sample HLA panel of the main atlas is not a coincidental hit (Supplementary §S1). Three families in the atlas (GPCR coupling, protease, kinase) retain the qualitative crosscutting verdict from the companion paper rather than a recomputed quantitative index, because their source alignments were not re-derived here. And the second dimension remains a post-hoc structural read; making locality itself predictable from sequence — for instance from the concentration of the anomaly distribution, which we found too confounded with the score to serve as an independent predictor — is the natural next problem.

The same crosscutting failure mode recurs beyond the seven families: on 43 sarbecovirus receptor-binding domains grouped by ACE2 usage — a grouping that nearly coincides with viral clade — the raw specificity signal is dominated by the clade confound, and only three receptor-binding-motif positions survive a phylogeny-aware null (Fisher $p = 0.044$; a weak-to-moderate confirmation, detailed in Supplementary Information SI-B).

Relation to the companion preprints. Companion preprint P1 introduced the conservation–divergence spectrum, established its geometry and information-theoretic reading, and demonstrated that it recovers known specificity determinants across three superfamilies. This paper asks the complementary question — *when does it work, and can you know before you look* — and answers it with a criterion that is prospectively falsifiable. The two papers share no primary result: P1's families (GPCR coupling, serine proteases, protein kinases) establish the instrument; this paper's added families (HLA, nuclear-receptor DBD and LBD, opsin tuning, aminoacyl-tRNA synthetases) map its domain of validity. Together they constitute an instrument and its operating manual. A third companion preprint, P3, takes the instrument to a position-by-position confrontation with a protein language model on GPCR coupling, and supplies the permutation-null machinery — shared with this paper — that shows why the family-specific thresholds mapped here cannot transfer. All three are posted together (Research Square, 2026; cross-referencing DOIs to be added on posting).

Methods

Spectra. For each family, sequences were aligned to the relevant Pfam profile HMM to obtain structurally equivalent match-state columns (full derivation of the conservation–divergence spectrum in companion preprint P1). For each column we computed within-class conservation as $1 - H_{\text{within}}/\log_2 20$ (mean over classes of the within-class residue entropy) and between-class divergence as the Jensen–Shannon divergence among the per-class residue distributions, normalized by $\log_2(\text{number of classes})$. The anomaly score is the residual of between-class divergence against a conservation-binned trend, isolating positions more divergent than their conservation predicts. Effective sequence counts use Henikoff position-based weights [5]; families were subsampled for domain/taxonomic balance before weighting.

Crosscutting index (dimension 1). Computed from functional labels and taxonomy only, never from the alignment columns under test. For cross-species families, we parsed the NCBI taxonomic lineage of every organism and recorded, for each functional label, the deepest rank (domain > kingdom > phylum > class > order) at which it recurs in ≥ 2 independent lineages (recurrence depth), together with the normalized mutual information between label and taxon rank. For within-species paralog families, where species taxonomy is constant, we substituted the normalized mutual information between the label and a gene-duplication tree built from the sequences themselves, and required the functional label to derive from an assay independent of the sequence (peptide-elution motif for HLA; endogenous hormone for NR-LBD) to avoid circularity.

Prospective tests. For each prospective family we froze, in a version-controlled pre-registration file before scoring: the predicted verdict, the crosscutting basis, the gold-standard residues (each verified by identity against a named reference UniProt sequence and mapped to Pfam match states), the pass/fail rule, and a stop-loss condition. aaRS: subclass Ia (PF00133), three cognates

(Ile/Leu/Val), gold standard the isoleucyl amino-acid pocket verified against *T. thermophilus* P56690, structural validation against PDB 1FFY. NR-LBD: PF00104, eight ligand classes, gold standard the estrogen-receptor- α estradiol pocket verified against P03372, structural validation against PDB 1ERE. Recall counts gold-standard residues among the top 15% / top 5% of anomaly; the gatekeeping test (aaRS) permutes labels within each domain 500 $\times$ and applies Benjamini–Hochberg correction; structural enrichment correlates anomaly with distance to the ligand/substrate in the holo structure and tests a permutation null on top-percentile mean distance.

Benchmark. Mutual information, SDPpred-style cumulative per-class relative entropy versus the overall distribution, and an evolutionary-trace-style product of mean within-class conservation and number of distinct class modes, each computed on the same aligned matrices and scored by the same gold-standard recall as the anomaly score.

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 and the per-position master tables are 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 P3, 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. Mirny LA, Gelfand MS. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. *J Mol Biol.* 2002;321(1):7–20. doi:10.1016/S0022-2836(02)00587-9
2. Hannenhalli SS, Russell RB. Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol.* 2000;303(1):61–76. doi:10.1006/jmbi.2000.4036
3. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol.* 1996;257(2):342–358. doi:10.1006/jmbi.1996.0167

4. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*. 1999;286(5438):295–299. doi:10.1126/science.286.5438.295
5. Henikoff S, Henikoff JG. Position-based sequence weights. *J Mol Biol*. 1994;243(4):574–578. doi:10.1016/0022-2836(94)90032-9
6. Sette A, Sidney J. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics*. 1999;50(3-4):201–212. doi:10.1007/s002510050594
7. Eriani G, Delarue M, Poch O, Gangloff J, Moras D. Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature*. 1990;347(6289):203–206. doi:10.1038/347203a0
8. Yokoyama S. Molecular evolution of vertebrate visual pigments. *Prog Retin Eye Res*. 2000;19(4):385–419. doi:10.1016/S1350-9462(00)00002-1
9. Shen H. A conservation–divergence spectrum reads subtype specificity from sequence as a sparse, distributed, probabilistic code across protein superfamilies. *Research Square*. 2026. doi:10.21203/rs.3.rs-10633466/v1

Supplementary

S1 — The HLA hit is robust to sample size and to the grouping label

The main-text HLA spectrum rests on the small HLA-B allele set of the calibration atlas, inviting the concern that its four-corner recovery is a small-sample coincidence. We addressed this directly by expanding HLA-B from IPD-IMGT/HLA (4,138 complete, 2-field-distinct $\alpha 1/\alpha 2$ sequences, MAFFT-aligned to the same 341 mature-position coordinate frame as the atlas) and re-running the conservation–divergence null analysis under two orthogonal manipulations.

First, holding the grouping fixed, we walked a sequence-weighting ladder — uniform, Henikoff position-based, 98%-identity de-duplication, real AlleleFrequencies.net population-frequency weighting, and supertype-consensus collapse. All five schemes returned null_explains in [1.00, 1.07], matching the other cross-species families and confirming the at-scale recovery is not a weighting or pseudoreplication artifact.

Second, and directly relevant to the circularity argument, we compared two labels on an identical set of 63 HLA-B alleles carrying measured eluted-peptide motifs (MHC Motif Atlas, ≥ 20 nine-mers per allele). The **independent** label — the P2 anchor-residue chemistry read from the eluted peptides, not derived from the sequence columns under test — gives null_explains = 0.957 (95% bootstrap CI over positions 0.881–1.058): a hit at N = 63. The **circular** label — the sequence-defined pocket supertype — gives 0.766 (CI 0.689–0.834) on the same alleles, reproducing the anomalous sub-null signature. A trivial-convergence control (subsampling the circular-label set from 5 to 240 alleles per group, three seeds) shows the circular anomaly relaxes toward 1.0 only as N grows (range 0.86–1.18, never approaching the 0.78 of the atlas panel), whereas the independent label is already at the null at small N. The two manipulations thus isolate two separable contributors to a spuriously low

null_explains — a circular grouping label and small sample size — and show that removing either one restores the hit.

Scope. The independent (non-circular) label exists only for assayed alleles (63 HLA-B here); this is a legitimate expansion of the *independent-label* set, not a thousands-allele independent analysis — the 4,138-allele arm necessarily uses the sequence-defined label and serves only as the sample-size / weighting-robustness control. These runs use the permutation-null null_explains statistic, which corroborates the HLA hit's robustness but does not replace the main-text HLA spectrum or its p and recall figures. frac_above_null is sample-size-dependent and is therefore not compared across N; null_explains with a bootstrap CI over positions is the reported statistic.

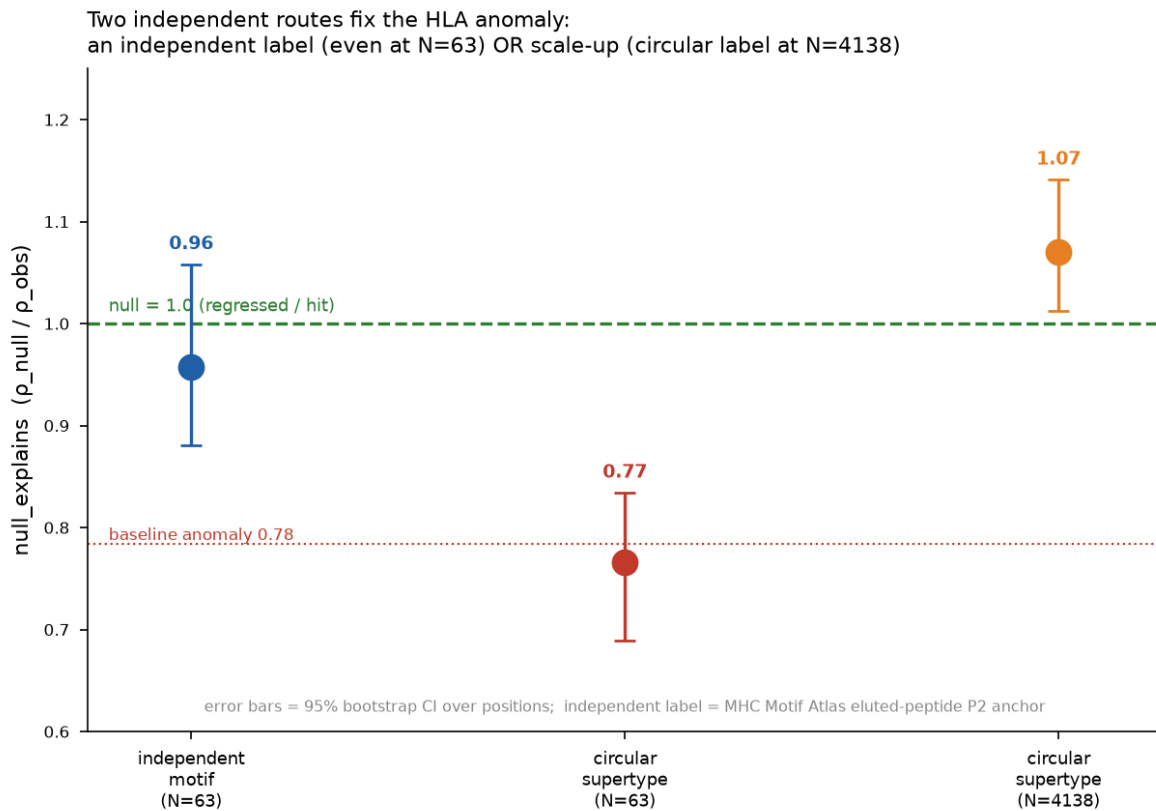


Fig S1. null_explains under {independent motif, N = 63}, {circular supertype, N = 63}, and {circular supertype, N = 4,138}, with 95% bootstrap CIs; the baseline atlas anomaly (0.78) and null = 1.0 are marked. Either an independent label at small N or scale-up of the circular label restores the hit.

Data and code. All master tables, alignments, pre-registration files, and figure-generation code are provided as supplementary datasets (see companion data manifest). Supplementary §S1 data: hlaB_independent_label.npz (63 assayed alleles + independent motif labels), independent_vs_circular_label.csv, persite_independent_P2motif.csv, rung_summary_final.csv and regression_criterion_bootstrap.csv (five-tier weighting ladder + bootstrap CIs). Sources: MHC Motif Atlas class I eluted peptides (nine-mers, ≥20/allele); IPD-IMGT/HLA B_prot (Latest).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [FigS1.png](#)
- [dataavailabilitystatement.txt](#)
- [P2combined.pdf](#)