A conservation–divergence spectrum reads subtype specificity from sequence as a sparse, distributed, probabilistic code across protein superfamilies

Huazhang Shen

Independent Researcher

Correspondence: 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Preprint, not peer reviewed.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: this paper (P1) introduces the conservation–divergence spectrum, 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. Proteins in a shared-fold family use highly conserved positions to encode functions common to the whole family, and a smaller set of positions to encode subtype-specific function. How that second, specificity-encoding layer is written into sequence — and whether it obeys any general organizing principle — remains only partly understood. Here we place every structurally equivalent position of a protein family on a plane whose axes are within-class conservation and between-class divergence, and show that the resulting cloud is not an arbitrary scatter but a trend band constrained by an information-theoretic feasibility boundary. The two axes are negatively correlated (Spearman $\rho = -0.64$), but much of this negative correlation is generated by a matched permutation-null floor (companion preprint P3); the biologically specific signal is the position-wise excess above that null, not the raw trend. The plane's four corners correspond to four selection regimes, and its sparse conserved-yet-divergent corner isolates candidate specificity-determining positions (SDPs). Applied to G-protein-coupling specificity across 219 class-A G-protein-coupled receptors (GPCRs), the framework recovers a coupling code that is distributed over ~12 mostly buried positions rather than localized to a binding interface, is consistent with a steric rather than electrostatic encoding through side-chain volume, and is consistent with a phenomenological Boltzmann-competition interpretation among G proteins — a probabilistic rather than a deterministic code — though we do not measure binding free energies directly. The same code shows a spectrum of evolutionary ages, from anchors frozen for ~450 My to positions still turning over across vertebrate lineages. Because the method's pattern is consistent with the sequence shadow of a general physical law — non-covalent recognition set by free-energy differences that Boltzmann statistics amplify, a phenomenological interpretation rather than a direct energetic measurement — it transfers across superfamilies: from sequence alone and with no family-specific input, it recovers the textbook S1 specificity determinant of serine proteases (position 189) and the catalytic-loop Tyr-versus-Ser/Thr determinants of protein kinases. We position the framework as a sequence-level ranking lens — a fast, general prospecting instrument that reorders a family's positions by their probability

of carrying specificity — not as a per-position decision procedure. The verdict on any single candidate remains with structure, energetics or experiment.

Keywords: conservation-divergence spectrum; specificity-determining positions; GPCR coupling; protein superfamily; information theory; random forest classifier; molecular evolution; sequence-based prediction

Author Summary

A protein family that shares one fold still splits into subtypes that do different things — different receptors couple to different signaling proteins, related enzymes cut different substrates. The positions that keep the shared machinery working are easy to find because they are conserved. The positions that make each subtype different are harder: they are few, scattered, and individually weak. We show that plotting every position of a family on two axes — how conserved it is within each subtype, and how different it is between subtypes — organizes this problem. The positions that matter for specificity sit in one sparse corner, and where that corner falls is bounded by a hard information-theoretic limit. Using this plane on G-protein-coupled receptors, we find that the "address label" that decides which signaling protein a receptor talks to is written probabilistically across roughly a dozen buried positions, each nudging the odds by a fraction of a thermal energy unit, rather than being decided by any single switch. Because the plane measures the sequence trace of a universal physical rule, the same code — run blind — rediscovers well-established specificity determinants in two unrelated enzyme families. Throughout, we treat the method as a way to *rank* which positions are worth investigating, not as a way to *decide* what any one position does; that decision belongs to structure and experiment.

Introduction

Proteins that share a fold and a core mechanism nonetheless diversify into functional subtypes: one G-protein-coupled receptor activates Gs and another Gi, one serine protease cuts after basic residues and another after aromatics, one kinase phosphorylates serine and another tyrosine. The information that distinguishes subtypes must be written somewhere in the sequence, superimposed on the information that maintains the shared fold and mechanism. Separating these two layers — the positions that encode what the family has in common from the positions that encode what makes each subtype different — is a recurring problem in molecular evolution, protein engineering and functional annotation.

The two layers leave opposite statistical signatures. Positions that maintain shared function are under uniform purifying selection and are therefore conserved across the whole family; they are straightforward to find by conservation alone. Positions that encode subtype-specific function are harder. They are conserved within each subtype (each subtype needs its own residue there) but differ between subtypes, and there are few of them. A large methodological literature has developed to find these specificity-determining positions — from early conservation-and-divergence scores and the mutual-information and statistical-coupling analyses that detect coevolving networks [1,3], through direct-coupling analysis [2] and, more

recently, the coevolutionary signals that deep models exploit. Most of these methods return a ranked list of candidate positions. Fewer ask what organizing principle, if any, governs where specificity positions can lie, how the specificity signal is distributed in space, how strong it is in physical units, or whether the same principle holds across superfamilies.

Here we take a deliberately simple and general view. We place every structurally equivalent position of a protein family on a two-dimensional plane whose axes are within-class conservation and between-class divergence, and we ask what the geometry of the resulting cloud reveals. The central observations are that (i) the occupied region is bounded by a hard information-theoretic ceiling, and much of the two axes' negative correlation is generated by a matched permutation-null floor rather than by a biological trade-off; the signal of interest is the position-wise excess above that null, quantified across four families in companion preprint P3; (ii) the sparse conserved-yet-divergent corner isolates candidate specificity-determining positions, which we treat throughout as high-probability candidates from a ranking instrument rather than as decided assignments; (iii) for the well-characterized case of GPCR–G-protein coupling, the specificity code is distributed across roughly a dozen mostly buried positions, is consistent with a steric rather than electrostatic encoding, carries only a small modelled energetic bias per position, and spans a wide range of evolutionary ages; (iv) this pattern is consistent with a phenomenological Boltzmann-competition interpretation of recognition among binding free energies — a probabilistic rather than a deterministic code — though we do not measure binding free energies directly; and (v) because the underlying instrument is general, the same plane recovers known specificity determinants in unrelated enzyme superfamilies when run blind.

Our contribution is methodological. The individual physical and statistical ideas we draw on are not new: that molecular recognition is governed by free-energy differences that Boltzmann statistics amplify; that specificity often rests on marginal energetic margins and promiscuity is common [5]; that binding energies are approximately additive and statistically distributed [4]; and that specificity in binding can be treated as a general statistical-mechanical problem [6]. What we add is a projection of these ideas onto a single, family-agnostic sequence/evolution/structure coordinate system, and the demonstration that reading a family's positions in this coordinate system reproduces, unifies and extends what is separately known. We frame the plane as a fast, general prospecting instrument that reorders a family's positions by their probability of carrying specificity — useful precisely because it is cheap, requires only an alignment and a grouping, and transfers across families — while leaving the verdict on any individual position to structure, energetics and experiment.

We develop the framework on class-A GPCR G-protein coupling because it is exceptionally well characterized — hundreds of receptors with measured coupling [8–15,22], a shared generic-numbering scheme, and abundant structural and mutagenesis data against which sequence-level predictions can be checked — before testing its transfer to serine proteases and protein kinases. The project began with a narrow question about serotonin-receptor coupling and broadened, once the geometry recurred, into the family-agnostic framework presented here.

Results

1. A conservation–divergence spectrum and its geometry

We analyze a protein family through a single plane. Given a multiple-sequence alignment whose columns are structurally equivalent positions — for GPCRs, the Ballesteros–Weinstein/GPCRdb generic numbering of the seven transmembrane helices; for enzyme families, the match states of a Pfam profile HMM — and a functional grouping of the sequences, we assign each position two coordinates. **Within-class conservation** measures how strongly a position is fixed inside each functional class (the strength of purifying selection acting on it), computed as 1 − H_within / $\log_2 20$ from the class-weighted mean within-class Shannon entropy. **Between-class divergence** measures how differently the classes use that position, computed as the mean pairwise Jensen–Shannon divergence (JSD) between class-specific residue distributions. Every position becomes one point on the plane (Fig 1B).

The two axes are not independent. Across all 212 aligned positions of 219 human class-A GPCRs grouped by primary G-protein coupling (Gs, Gi/o, Gq/11), conservation and divergence are strongly negatively correlated (Spearman $\rho = -0.64$, $p \approx 1.5 \times 10^{-25}$). This negative trend is largely generated by a matched permutation-null floor that rises as conservation falls, not by a direct biological trade-off between two selection pressures — companion preprint P3 runs an independent divergence computation on this same family and finds a very similar observed trend ($\rho \approx -0.61$) together with a null trend that is stronger still ($\rho_{null} \approx -0.90$); the two papers' divergence pipelines are not numerically identical; each should be read from its own paper. The feasibility boundary is information-theoretic; the signal of interest is the position-wise excess above that null, quantified position-by-position across four families in companion preprint P3 (Research Square, 2026; DOI to be added on posting).

The plane's four corners are interpretable as four selection regimes (Fig 1A); we refer to this layout as the framework's four-corner geometry. The lower-right corner (high conservation, low divergence) holds positions under uniform strong purifying selection — the allosteric-switch backbone shared by all class-A receptors (D2.50, the DRY motif at R3.50, P6.50, the NPxxY motif at 7.50), each frozen to the same residue in every coupling class. The upper-right corner (high conservation and high divergence) holds positions under class-specific purifying selection — each class fixes a different residue and holds it — and these are the candidate SDPs. The upper-left corner (low conservation, high divergence) holds fast-evolving, class-unstable positions whose apparent divergence is soft. The lower-left corner is nearly empty, and necessarily so: a position variable within every class converges to the background residue distribution in each class, so the between-class JSD collapses toward zero. "Highly variable yet strongly divergent" is close to a mathematical impossibility, not a biological absence.

The upper-right corner's sparseness is likewise structural. To enter it a position must *simultaneously* be functionally important (held within each class) and functionally differentiated (needing different residues between classes); the double requirement is intrinsically stringent. The cleanest operational definition of a specificity anchor follows directly: it is a position that **breaks the conservation–divergence trend** — one that "should" be low-divergence given its high conservation but is not, signaling active class-

specific selection maintaining the difference. Three positions meet the strictest form of this test (upper-right corner ∩ bootstrap selection frequency > 80% ∩ top-8 anchor weight): 3.21, 4.59 and 5.65 (circled in Fig 1B).

Finally, the plane has a hard ceiling. For a three-class grouping the between-class JSD cannot exceed min(log23, log220 · conservation): it is bounded above by log23 ≈ 1.58 bits (the maximum divergence of three distributions) and, at low conservation, by the residual entropy budget itself. This feasibility boundary is a concave arc (Fig 1C). The striking observation is how far the data sit *below* it. The maximum observed between-class divergence is 0.32 bits — only ~20% of the way to the log23 ceiling — and the point cloud is statistically indistinguishable from a straight line (quadratic fit RMSE 0.0462 vs linear 0.0463, no improvement); the faint visual "arc" is produced by the handful of anchors in the upper-right corner. No position in the family approaches the deterministic-code limit where each class would lock a different residue. Even the strongest anchor, 4.59 (modal Leu in Gs, Pro in Gi/o and Gq/11), encodes only a statistical preference. The coupling code is probabilistic and redundant, not look-up-table-like — a conclusion the remaining sections develop from independent evidence.

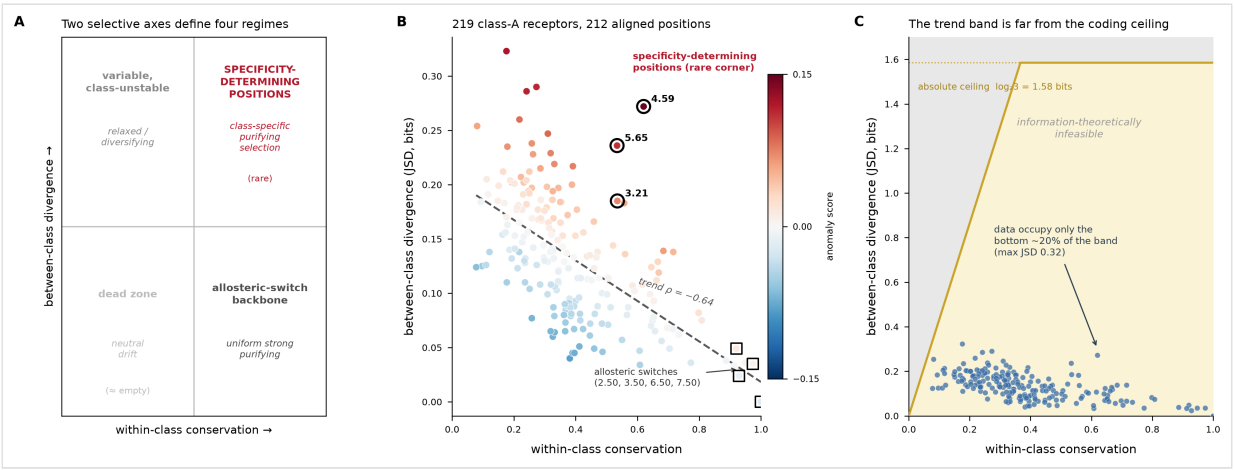

Fig 1. A conservation–divergence spectrum and its information-theoretic geometry. (A) The two selective axes — within-class conservation and between-class divergence — partition positions into four regimes: an allosteric-switch backbone (conserved, non-divergent), candidate specificity-determining positions (conserved yet divergent, rare), fast-evolving class-unstable positions (variable, divergent), and a near-empty neutral-drift corner. **(B)** The plane for 219 human class-A GPCRs across 212 generic-numbered positions, grouped by primary G-protein coupling; points colored by specificity-anchor weight (bootstrap-stabilized anchor score). The dashed line is the conservation–divergence trend (Spearman ρ = −0.64); red circles mark the three triple-criterion core anchors (3.21, 4.59, 5.65); blue squares mark canonical allosteric switches. **(C)** The same points against the information-theoretic feasibility boundary min(log23, log220·conservation). Data occupy only the bottom ~20% of the feasible band (maximum JSD 0.32 bits vs the log23 = 1.58-bit ceiling), indicating a probabilistic rather than a deterministic look-up code.

2. The signal is real, and it depends on the question asked

Two objections must be met before the upper-right corner can be read as biology. The first is that "conserved-yet-divergent" might simply be what conservation buys: more conserved positions could look more divergent for statistical reasons alone. The second is that any residue–coupling association could be a phylogenetic artifact — closely related receptors share both residues and coupling, so a position can appear coupling-informative merely by tracking subfamily membership.

We address the first by defining an **anomaly score** as each position's between-class divergence minus the value predicted by the conservation trend, explicitly subtracting the "conserved therefore divergent" baseline (Fig 2A). Positions with positive anomaly diverge *more* than their conservation predicts — the operational signature of active class-specific selection. We address the second with a phylogenetically bias-corrected gatekeeping test. For each position we compute the conditional mutual information I(residue; coupling | subfamily) and compare it to a null generated by permuting coupling labels *within* each subfamily, so that the observed and null statistics share the same finite-sample estimation bias and it cancels on subtraction (Fig 2B). After Benjamini–Hochberg correction, **141 of 212 positions retain coupling information beyond subfamily structure at FDR < 0.05**, and all top-ranked anchors — including 4.59, 5.65 and 3.21 — survive. The coupling signal is therefore not a phylogenetic shadow; a majority of positions carry residue–coupling association that within-subfamily permutation cannot explain away.

The framework's second axis is defined by the functional grouping, and this is a feature rather than a limitation: changing the grouping asks a different question, and the specificity corner moves accordingly. Holding the same set of positions fixed (209 of the 212 with sufficient coverage across all four groupings) and regrouping the 219 receptors four ways — by G-protein coupling, by ligand type, by G-protein-versus-arrestin bias, and by phylogenetic subfamily — produces four different sets of high-divergence positions (Fig 2, bottom row). The three grouping logics that correspond to distinct biological questions localize to distinct regions of the receptor. **Ligand-specific** positions concentrate in the mid-membrane orthosteric pocket (3.32, 3.36, 6.52, 7.43) — where the ligand is bound. **Coupling-specific** positions are distributed along the helices rather than clustered at the intracellular G-protein interface; the top-divergence positions include 3.29, 4.60 and 5.66, several of which sit toward the extracellular half. **Bias-specific** positions fall toward the extracellular ends (6.58, 6.61, 1.35). The phylogenetic grouping, by contrast, marks positions that simply track subfamily identity and carries no localized functional signal. That the y-axis is contingent on the grouping is exactly why the plane is a general instrument: it does not discover function on its own, but given a functional partition it reports which positions carry that partition's specificity, and where they sit.

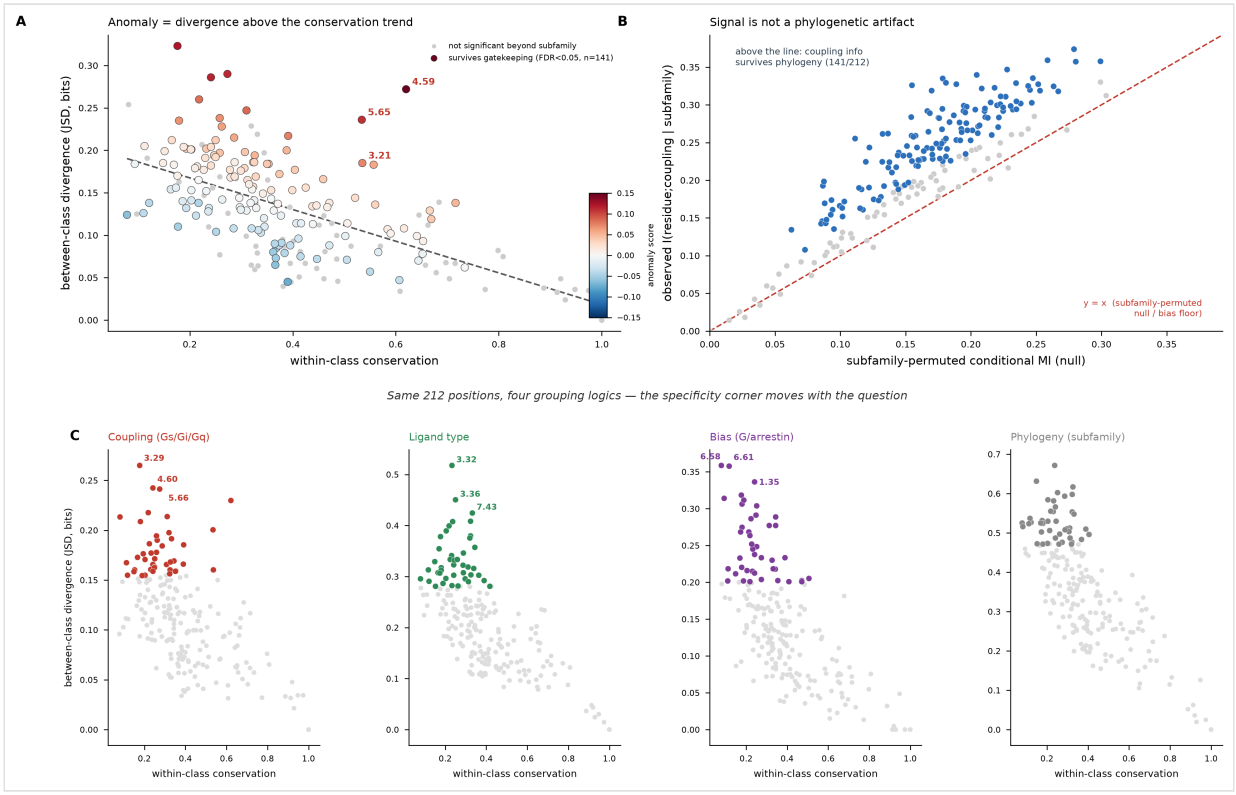

Fig 2. The specificity signal survives phylogenetic control and is defined by the grouping. (A) The conservation–divergence plane colored by anomaly score (divergence above the conservation trend); grey points carry coupling information not significant beyond subfamily structure, colored points survive gatekeeping. **(B)** Bias-corrected phylogenetic gatekeeping: observed conditional mutual information I(residue; coupling | subfamily) versus a within-subfamily-permuted null. Points above the identity line retain coupling information beyond subfamily structure; 141 of 212 positions pass at FDR < 0.05. **(C, bottom row)** The same 209 positions (of 212, those with sufficient coverage across all four groupings) regrouped four ways. High-divergence positions (top quintile, colored) localize differently for each question: ligand type → central orthosteric pocket; coupling → distributed along helices; G/arrestin bias → extracellular ends; phylogeny → no localized signal. Selected top positions are labeled with generic numbers.

The next three sections turn to a single, deeply characterized superfamily — GPCR coupling — to ask what the specificity signal identified above physically is (§3), how old it is (§4), and what it predicts downstream (§5); readers interested primarily in the transfer of the general method beyond GPCRs can proceed directly to §6.

3. Coupling specificity is a distributed, buried shape-modulator network

If the upper-right corner marks real specificity positions, what do they physically do? A random-forest classifier trained on residue identity at these positions predicts a receptor's primary G-protein coupling (Gs, Gi/o, Gq/11) at 82% overall / 0.78 balanced accuracy under leave-one-out cross-validation (permutation p = 0.032), well above the 53% majority baseline (Fig 3A,B). Two features of *how* it succeeds define the code. First, a 53-position **interface** set outperforms the 14-position orthosteric **pocket** set (0.82 vs 0.76 overall accuracy), and

adding the pocket to the interface barely improves on the interface alone (0.83) — coupling specificity lives at the intracellular interface region, not in the ligand pocket. Second, no single position carries the signal; the classifier needs roughly a dozen positions distributed across TM5, TM6 and TM7, consistent with the plane's finding that coupling divergence is spread along the helices (§2).

The physical encoding is consistent with a steric, rather than electrostatic, mechanism. Comparing the modal side-chain properties of each position across coupling classes, specificity tracks side-chain volume and hydrophobicity, while charge is essentially invariant between classes (Fig 3C): the between-class range of mean charge is near zero at every position, whereas volume ranges up to ~30 Å³. This residue-property association does not by itself establish the underlying biophysical mechanism; we report it as a phenomenological pattern. The strongest anchors illustrate the pattern — 5.65 switches between a small residue in Gs (Ala) and larger residues in Gi/o and Gq/11 (~32 Å³ range), and 4.59 between a large hydrophobic residue in Gs (Leu, ~143 Å³) and smaller residues elsewhere. Of the specificity positions we could assign structurally, ~77% are hydrophobic.

Most of these positions do not touch the G protein. Measuring each anchor's minimum distance to Gα in the 5-HT2A–mini-Gαq complex, only **3 make direct contact (<5 Å; 3.54, 6.37, 5.65)**, 3 lie in the interface shell (5–8 Å), and **16 of 22 are buried in the transmembrane bundle (>8 Å)** — including the highest-scoring anchor, 4.59, at 31 Å. Anchor strength is uncorrelated with distance to Gα (Spearman $\rho = -0.03$, $p = 0.91$; Fig 3D): buried positions carry as much specificity information as contacting ones. Discriminative power is redundantly distributed. Of the 26 top-ranked specificity positions, the 3 direct-contact positions alone classify only marginally above baseline (0.58 overall accuracy) and the 3 shell positions alone reach 0.55, but the 6 interface positions (contact + shell) together reach 0.66; excluding those 6, the remaining 20 non-interface positions alone still reach 0.69. (Distances here are the 22 anchors we could locate in the complex structure — 3 contact, 3 shell, 16 buried; the discrimination subsets are drawn from the 26-position top-ranked set.) No single region is necessary.

Together these observations describe a distributed, buried shape-modulator network. Specificity is not a lock-and-key readout at the contact surface; it is encoded in the shape of the intracellular cavity into which the Gα α5 helix inserts, set collectively by a ring of mostly buried hydrophobic positions modulating side-chain packing. This clarifies the division of labor that runs through the whole system: the conserved allosteric switches (D2.50, DRY, PIF, NPxxY) are near-perfectly conserved yet carry essentially no coupling-discriminative information (balanced accuracy ≈ 0.56, indistinguishable from baseline), because the action they encode — the TM6 outward swing of activation [16,17] — is shared by all class-A receptors regardless of which G protein is engaged, consistent with a companion analysis of allostery between distant structural regions in GPCR:G-protein coupling [23]. The conserved layer is the pipe that carries the signal; the divergent layer is the valve that sets where it goes. A conserved position cannot discriminate coupling precisely because it is doing a job common to every subtype — the structural meaning of "conservation ≠ discriminative power."

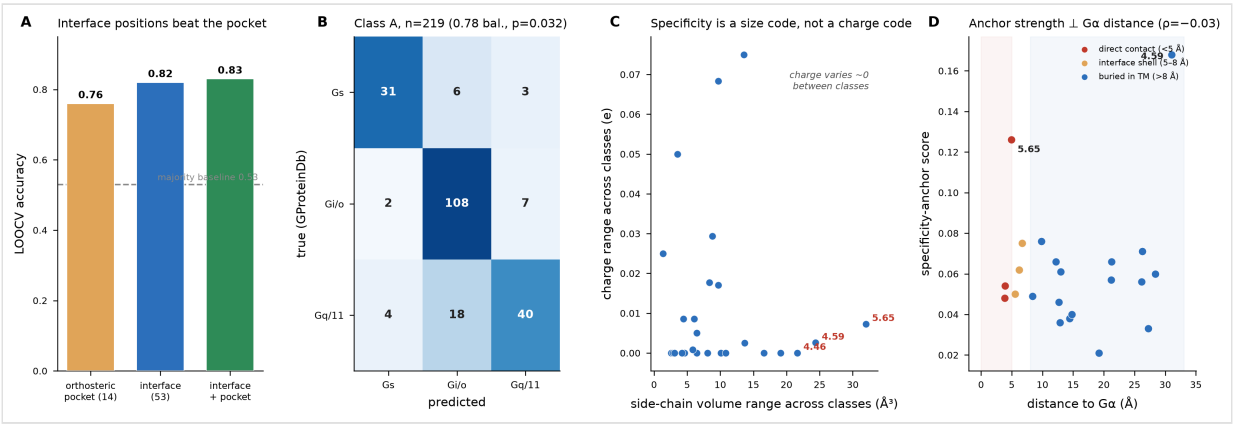

Fig 3. Coupling specificity is a distributed, buried, steric code. (A) Leave-one-out overall accuracy of a random-forest coupling classifier using different position sets; the 53-position interface set (0.82) outperforms the 14-position orthosteric pocket (0.76), and the majority baseline is 0.53. **(B)** Confusion matrix for the interface classifier (n = 219; 82% overall / 0.78 balanced accuracy, permutation p = 0.032). **(C)** Between-class range of modal side-chain volume versus modal charge, per specificity position: volume varies by up to ~30 Å³ while charge is near-invariant — a size code, not a charge code. **(D)** Specificity-anchor score versus minimum distance to Gα (5-HT2A–mini-Gαq): score is uncorrelated with distance (ρ = −0.03), and 16 of 22 assignable anchors are buried (>8 Å), including the top anchor 4.59.

4. Specificity anchors have an evolutionary age spectrum

The conservation–divergence plane is a static snapshot. Adding an evolutionary axis reveals that positions occupying the same corner today reached it on very different timescales. We quantify positional evolution three ways, all consistent. **Cross-species drift** — the within-receptor entropy of a position across orthologs — is anticorrelated with conservation (Spearman ρ = −0.54; a precheck confirming the measure is not tautological) and cleanly separates the plane: the allosteric switches are fully frozen (drift = 0), the core anchors low, other positions spread (Fig 4A). **Class-mode turnover** across four evolutionary layers (primate → mammal → bird/reptile → fish, ~450 My; coupling labels propagated along orthologs with 99% self-consistency) is anticorrelated with anchor strength (ρ = −0.35, $p = 2.7 \times 10^{-7}$; Fig 4B): stronger anchors change their class-specific residues less often. **Ancestral reconstruction** by Fitch parsimony on a 235-sequence tree counts the number of substitution events each position underwent, and this too falls with anchor strength (ρ = −0.41; Fig 4C).

The three measures converge on a graded picture. The allosteric switches D2.50 and R3.50 register **zero** ancestral changes — frozen for roughly 450 My. The core anchors are stable but not frozen: 3.21 = 6, 5.65 = 7, and the top-scoring anchor 4.59 = 8 events, with zero four-layer turnover at 4.59 and 3.21. At the other extreme, the edge anchor **5.66 has 29 ancestral changes, 100% turnover across all four layers, the highest cross-species drift and the lowest bootstrap stability of the core set** — it has been recruited into the coupling code repeatedly and independently in different lineages rather than inherited from a common

ancestor. The same corner of the plane therefore holds positions of very different evolutionary age: some locked for hundreds of millions of years, others still in motion.

Position 5.66 rewards a closer look because it is where our sequence-evolution signal and the independent mutagenesis literature converge (Fig 4D). Its residue identity is class-graded, not class-determined: in the 219 receptors its modal residue is Arg in Gs (42%), a mix of Arg/Lys/Leu in Gi/o, and — distinctively — Trp in Gq/11 (21%). Trp is merely enriched in Gq/11, far from fixed, so 5.66 is a probabilistic Gq-preference modulator rather than a Gq switch: it adds real signal to "is this Gq?" but is nearly silent for "Gs versus Gi/o." Independently, engineering studies of the β2-adrenergic and V1A receptors identify the TM5 residue at this position as a hotspot for the Gα peptide C-terminus, and shifting it toward Gq-type residues confers Gs→Gq promiscuity. The position our evolutionary analysis flags as "still moving" is the same one the mutagenesis field uses to reprogram coupling — a labile specificity knob, consistent across both lenses.

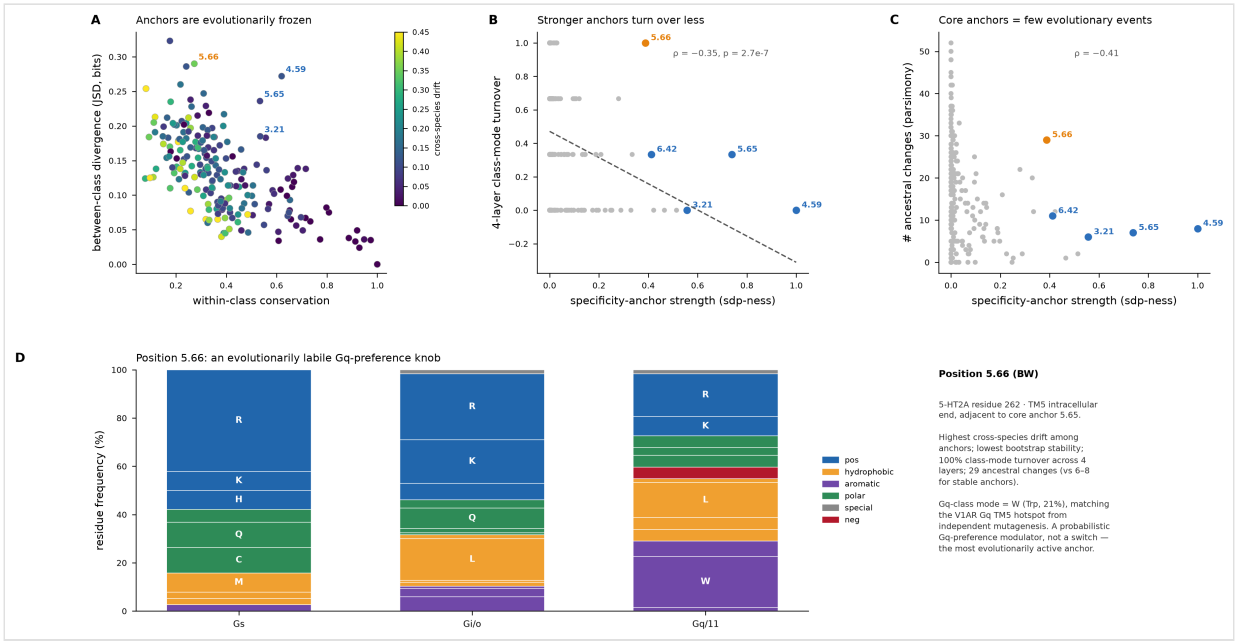

Fig 4. Specificity anchors span an evolutionary age spectrum. (A) The conservation–divergence plane colored by cross-species drift; allosteric switches and core anchors are frozen (pale), edge positions drift (dark). **(B)** Four-layer class-mode turnover versus specificity-anchor strength (ρ = −0.35): stronger anchors turn over less. **(C)** Number of ancestral substitution events (Fitch parsimony) versus anchor strength (ρ = −0.41); core anchors 3.21/5.65/4.59 register 6–8 events, edge anchor 5.66 registers 29. **(D)** Residue identity of position 5.66 by coupling class across 219 receptors (colored by residue property), with its multi-method profile: the Gq/11 mode is Trp (21%), matching an independently reported TM5 Gq hotspot.

5. The code is probabilistic: a Boltzmann competition

The sequence side is probabilistic (§1), so we asked whether the downstream output is a probability distribution too. A receptor faces a pool of G proteins competing for the same intracellular cavity, and the observed coupling probabilities are consistent with a Boltzmann-

competition interpretation over relative binding free energies, $P(G_i) \propto \exp(-\Delta G_i/kT)$ — a phenomenological reading of the data, not a direct thermodynamic measurement, since we do not measure binding free energies. Under this reading, at body temperature $kT \approx 0.6$ kcal/mol, so small modelled energy differences would be exponentially amplified into large probability differences: a $\Delta\Delta G$ of ~0.6 kcal/mol reading as a ~3-fold probability ratio, and ~1.4 kcal/mol as ~10-fold. This interpretation is consistent with the observations of the previous sections: it is consistent with why the code is probabilistic rather than a switch — the output resembling a Boltzmann softmax, so evolution need only tune cumulative side-chain volume to produce a modelled 1–3 kcal/mol of bias, never a large $\Delta\Delta G$; with why single positions discriminate weakly but a dozen together succeed — each anchor's modelled contribution of ~0.5 kcal/mol of $\Delta\Delta G$ adding linearly upstream of the exponential amplifier under this reading; with why the code tracks steric rather than electrostatic properties — steric complementarity supplying small, finely additive modelled $\Delta\Delta G$, whereas electrostatics is "harder" and carries desolvation penalties; and with why reprogramming shifts but does not flip coupling — changing a few modelled $\Delta\Delta G$ terms sliding the Boltzmann peak rather than toggling it. We present this as a phenomenological interpretation that organizes the data, not as a measured biophysical mechanism.

The prediction is directly testable in the coupling data. Converting quantitative coupling efficacies (log Emax/EC50) for 178 receptors into probability distributions and computing the coupling entropy of each yields a **continuous spectrum** (entropy 0–2 bits, median 0.59), not a bimodal split (Fig 5A): 44% of receptors couple exclusively, ~26% are promiscuous, and the intermediate range is continuously populated with no gap. Coupling promiscuity is the norm, not the exception. We initially expected balanced "transition-state" competition to be rare, by analogy with evolutionary transition states that are unstable and quickly left. The data overturn this: among 99 multi-coupling receptors with quantitative grades, the dominance margin ($p_1 - p_2$) is nearly flat, and near-degenerate coupling ($\Delta < 0.1$) is as common as any other regime — accounting for ~31% of them, not depleted (Fig 5B). The corrected reading is that "which G protein" is not a bistable switch but a flat post-activation competition: coupling both Gi and Gq is a stable, even actively selected phenotype (integrating multiple pathways), so there is no force pushing the system off $\Delta\Delta G \approx 0$. Receptors observed experimentally to compete between G proteins — the β2-adrenergic receptor's Gs↔Gi time-sharing, CCK1R's near-equal cavity for Gq and Gs, H2R's Gs/Gq promiscuity — all fall in this promiscuous zone (Fig 5C). The bistable behavior of the analogy belongs to the *activation* layer (short-lived intermediate conformations), not to the coupling layer — another facet of the conserved-switch / divergent-competition division of labor.

Finally, we tested whether the plane's ranked positions produce a consistent in silico response (Fig 5D). Training the coupling classifier and mutating 117 Gi/o receptors toward Gs modal residues, position by position in spectrum-recommended order, raises the model's mean P(Gs) from a baseline of 0.09 to 0.26 after 15 positions — roughly twice the shift obtained by mutating the same number of positions in random order (0.13). This shows the classifier's internal response to the ranked positions is systematically larger than to a random set of the same size; it demonstrates the model's sensitivity to the ranking, not the positions' causal or interventional effect on an actual receptor, which would require structural or experimental

validation. No receptor was flipped to a Gs majority (0% flip). We read this as consistent with a distributed probabilistic code: changing a set of positions slides the model's coupling preference rather than throwing a switch — consistent with reports that naturally reprogramming coupling (e.g. β2AR → Gq) requires remodeling an entire TM5 interface rather than a single site. This is an in silico prediction at the level of sequence fingerprints, not a structural or experimental validation, and we treat it as such (§Discussion).

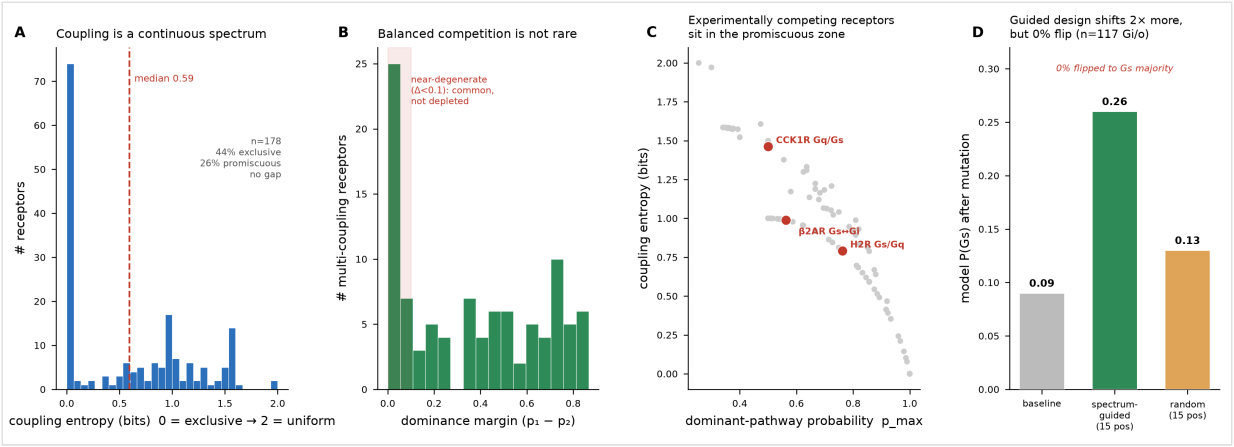

Fig 5. Coupling is consistent with a probabilistic Boltzmann-competition interpretation. (A) Coupling entropy across 178 receptors (0 = exclusive, 2 bits = uniform over four G proteins) is a continuous spectrum with median 0.59 and no bimodal gap. (B) Dominance margin ($p_1 - p_2$) among 99 quantitatively graded multi-coupling receptors (coupling entropy > 0.3 bits); near-degenerate competition ($\Delta < 0.1$, shaded; ~31%) is as common as any other regime. (C) Coupling entropy versus dominant-pathway probability; receptors experimentally reported to compete between G proteins (β2AR, CCK1R, H2R) fall in the promiscuous zone. (D) Reprogramming in silico: mutating 117 Gi/o receptors toward Gs modal residues in spectrum-guided order shifts model P(Gs) ~2× more than random order (0.26 vs 0.13 from 0.09 baseline), yet flips 0% of receptors — the code slides, it does not toggle.

6. The framework generalizes across superfamilies

If the conservation–divergence plane's pattern is consistent with the sequence shadow of a general physical law — non-covalent recognition set by additive free-energy differences amplified by Boltzmann statistics, a phenomenological interpretation rather than a direct energetic measurement — then it should not be specific to GPCRs. We tested transferability at two scales.

Within the GPCR superfamily, the coupling code learned on class A transfers to the other classes. For the 25 generic positions alignable between class A and classes B/C/F, a position's specificity divergence in class A predicts its divergence in the other classes (Spearman $\rho = 0.50$, $p = 0.012$; Fig 6A). The core anchors are shared: 5.65 and 4.59 carry coupling-specificity signal in both the class-A and the non-class-A alignments. The same structural positions are used to encode coupling specificity across receptor classes that diverged early in GPCR evolution.

The same transfer holds for the most sequence-divergent class A family. In the human olfactory receptors — regrouped by the ancestral Class I/Class II split (n = 62 vs 369) rather than by subfamily — 141 of 167 scored positions exceed the permutation null, and the Class I/II anomaly reproduces the subfamily-grouping anomaly at the same positions only partially (Spearman $\rho = 0.78$, $n = 167$): regrouping the same family along a different functional axis relocates part of the specificity signal, exactly as the question-dependence of §2 predicts. An independent structural check on the OR51E2 cryo-EM structure (PDB 8F76) confirms the intracellular coupling layer on 28 structure-verified positions. Olfactory receptors thus extend the within-superfamily transfer to a family that shares only the seven-helix fold with the class A receptors the method was built on. The per-family analyses and their structural caveats — including the corrected OR51E2 numbering — are given in Supplementary Information (SI-A); the companion P3 treats the olfactory boundary cases.

The stronger test is across unrelated superfamilies, run blind — no GPCR information, no family-specific tuning, only an alignment and a functional grouping. In **serine proteases** (trypsin vs chymotrypsin vs elastase; ~200 Pfam-aligned positions grouped by specificity class), the plane's top-anomaly corner recovers the textbook determinant of substrate specificity: **position 189** (chymotrypsin numbering), the base of the S1 pocket, with modal residues Asp/Ser/Ser across the three classes and anomaly rank 6 of 208 (Fig 6B) — the residue that structural enzymology established decades ago as the primary S1 specificity switch (Asp189 in trypsin creates the negatively charged pocket that binds Lys/Arg substrates). We did not tell the method this; it emerged from sequence and grouping alone. In **protein kinases** (Ser/Thr- vs Tyr-specificity; grouped by phosphoacceptor), the top-anomaly positions fall in the catalytic loop (subdomain VIb), the region that determines phosphoacceptor specificity (Fig 6C). The kinase signal is more distributed and less dominated by a single anchor than the protease signal — an honest difference between the families, not a forced match — but the specificity-carrying region is correctly identified.

These are two enzyme families, not a superfamily-wide survey, and we present them as proof-of-transfer rather than a claim of universality (§Discussion). But the blind recovery of two independently established, mechanistically different specificity determinants — an electrostatic pocket switch in proteases, a distributed catalytic-loop code in kinases — from the same conservation–divergence geometry indicates that the framework captures something general about how shared-fold families partition conserved from specificity-determining positions.

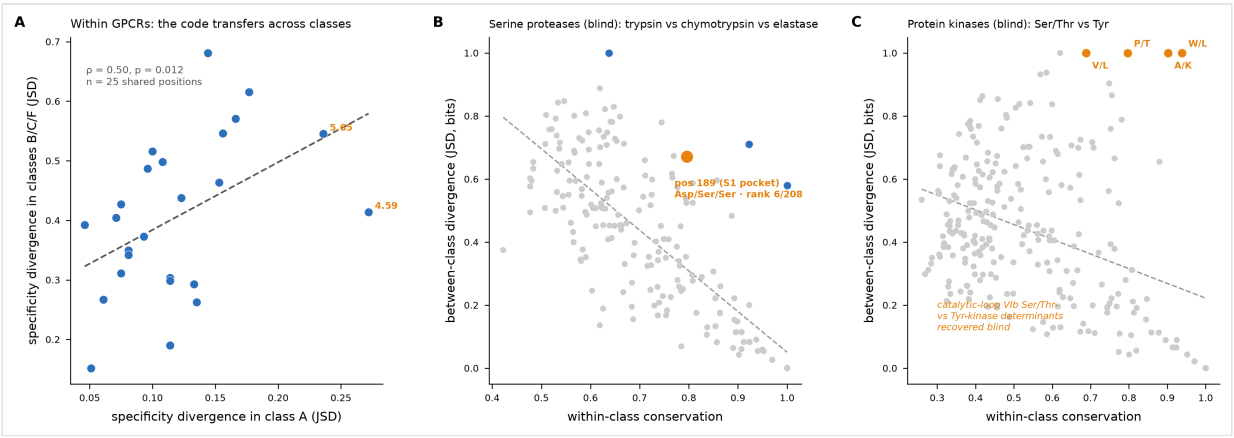

Fig 6. The framework transfers across classes and superfamilies. (A) Within GPCRs: for 25 positions alignable between class A and classes B/C/F, specificity divergence in class A predicts divergence in the other classes (ρ = 0.50, p = 0.012); core anchors 5.65 and 4.59 are shared. **(B)** Serine proteases, analyzed blind: the conservation–divergence plane's top-anomaly corner recovers position 189 (Asp/Ser/Ser; the S1-pocket specificity determinant) at anomaly rank 6 of 208. **(C)** Protein kinases, analyzed blind: top-anomaly positions fall in the catalytic loop (subdomain VIb), the Ser/Thr-versus-Tyr specificity region; the signal is more distributed than in proteases. Labeled residue pairs are the modal residues of the two kinase classes.

Discussion

We have described a single plane — within-class conservation against between-class divergence — and shown that its geometry organizes a family's positions into interpretable selection regimes, isolates a sparse corner of candidate specificity-determining positions, and, for GPCR coupling, exposes a code that is distributed, buried, steric, probabilistic and evolutionarily graded. Because the plane measures the sequence trace of a general physical law, it transfers across superfamilies. The value of the framework is that it ties observations usually made separately — specificity-position detection, spatial distribution, energetic magnitude, evolutionary age, cross-family generality — to one coordinate system and one physical reading.

What is and is not new here. The physical core — that recognition specificity can be described as a Boltzmann competition among binding free energies, that these energies are often marginal (a few kT), approximately additive and statistically distributed, and that promiscuity is therefore common — is established theory. Our contribution is to project that theory onto a sequence/evolution/structure coordinate system as a phenomenological interpretation and show the projection is internally consistent: the ~1 kT margins the theory would predict are consistent with a probabilistic coupling spectrum with no bimodal gap (Fig 5A); additivity is consistent with the linear accumulation of weak per-position discrimination into a competent distributed classifier (Fig 3); and the statistical view is consistent with the observation that guided reprogramming slides rather than flips coupling (Fig 5D). We claim

the mapping and its consistency with these observations, not a direct measurement of the underlying physics.

Boundaries. Several limits must be stated plainly. First, the specificity-position detector is a **signal-prospecting instrument, not a decision procedure**: it ranks positions by their probability of carrying specificity, and the effective sample size of the GPCR alignment (Meff ≈ 184 independent sequences) is enough to localize signal but not to certify any single position. Individual anchor assignments are hypotheses for structural or experimental follow-up. Second, the **G-protein/arrestin bias axis is not the coupling axis**: bias-divergent positions localize to the extracellular ends, distinct from the coupling code, and we do not claim the plane resolves biased signaling. Third, **cross-superfamily generality rests on two enzyme families** (serine proteases and protein kinases), which we present as proof-of-transfer, not as a survey; broader claims require systematic application across many families. Fourth, the **ancestral-reconstruction results are qualitative**: parsimony event counts on a single tree establish a rank order of evolutionary lability (frozen switches < core anchors < edge anchors), not calibrated divergence times, and the "~450 My" figure is the age of the vertebrate radiation spanned by the sequence set, not a dated substitution. Fifth, the reprogramming test is an **in silico** prediction on model-derived sequence fingerprints; it motivates experiments, it does not substitute for them.

Relation to structure prediction. Deep structure-prediction models [7] implicitly learn the coevolutionary statistics that encode both fold and specificity; our plane makes the specificity layer explicit and interpretable in physical units, which is complementary to, not competitive with, black-box prediction. We note the analogy only to place the work, and make no claim about improving or replacing such models.

Species sampling. The specificity plane is built from human paralogs, because its divergence axis is defined by a functional partition (coupling class) and it is the contrast between co-existing, differently-coupled receptors — not variation within one receptor's orthologs — that carries specificity information. Adding orthologs of a given receptor mostly injects near-redundant sequence without new coupling classes; the alignment's effective sample size (Meff ≈ 184 of 219 sequences) already reflects and down-weights this redundancy, so the negative conservation–divergence trend, the sparse conserved–divergent corner, and the far-below-ceiling occupancy are unlikely to move with broader paralog sampling. Species information does enter the framework, but on the age axis: cross-species drift, four-layer class-mode turnover, and the ancestral reconstruction all use ortholog sets and a 235-sequence vertebrate tree (Fig 4). The most tractable extension is therefore to deepen ortholog sampling per receptor, which would sharpen the lability rank order (which anchors are frozen in mammals yet labile in fish) and stabilize the ancestral inference. A more incisive but harder test would treat receptors whose primary coupling has itself shifted between species as natural experiments: if such a shift is accompanied by substitutions concentrated at the anchor positions, it would independently corroborate the probabilistic, low-margin picture that the in silico reprogramming only motivates. That test is gated not by sequence availability — orthologs are abundant — but by reliable cross-species coupling data, which at present are almost entirely human (GProteinDb and the 178 transduction-coefficient receptors alike); non-human quantitative coupling exists only as scattered pharmacological cases. Cross-species

validation of the specificity axis is thus realistically a case-by-case effort contingent on curating those few well-measured examples, whereas the evolutionary axis can be strengthened at scale today. A further reach — sampling invertebrate and basal-metazoan class-A receptors — could test whether the anchor set predates the diversification of the vertebrate coupling repertoire, i.e. whether the specificity coordinate is an ancient, reusable syntax.

Outlook. The framework is cheap and general: given an alignment and a functional grouping, it returns a ranked, spatially resolved, phenomenologically interpretable map of a family's specificity positions in seconds. Natural extensions are to run it systematically across Pfam families to test whether the conservation–divergence geometry — the null-corrected excess signature, the sparse conserved-divergent corner, the far-below-ceiling occupancy — is universal; to couple the ranked candidates to explicit ΔΔG calculations for the strongest anchors; and to use the plane as a prior for experimental specificity engineering — echoing calls for structure-informed, kinetically aware approaches to GPCR drug discovery [18,19] — where its message is consistent and useful: to change a subtype's specificity, edit the distributed set the plane identifies, and expect to shift the probability rather than throw a switch.

Methods

Sequences, alignment and functional grouping

Class-A GPCR analysis used human receptors with primary G-protein coupling annotated in GProteinDb [14], aligned on the Ballesteros–Weinstein [20] / GPCRdb [21] generic numbering of the seven transmembrane helices, giving 212 structurally equivalent positions across 219 receptors (Gs = 40, Gi/o = 117, Gq/11 = 62). The effective sample size of the alignment, estimated by down-weighting sequences above an identity threshold, is Meff ≈ 184 independent sequences; we use this to bound interpretation (signal localization, not per-position certification). Cross-species and ancestral analyses used ortholog sets and a 235-sequence tree spanning the vertebrate radiation. Serine-protease and protein-kinase analyses used Pfam profile-HMM match states as the position coordinate and phosphoacceptor / substrate-specificity class as the grouping; both were run with no GPCR-derived input.

Conservation and divergence axes

For each position, **within-class conservation** is $1 - \bar{H}_{within} / \log_2 20$, where $\bar{H}_{within}$ is the class-size-weighted mean of the within-class Shannon entropy of the residue distribution (gaps excluded), so conservation is 1 for a position fixed within every class and 0 for a uniform distribution. **Between-class divergence** is the mean pairwise Jensen–Shannon divergence (JSD, in bits) between the class-specific residue distributions. The **feasibility ceiling** is $\min(\log_2 K, \log_2 20 \cdot \text{conservation})$ for K classes (K = 3 for GPCR coupling; $\log_2 3 \approx 1.58$ bits), derived as the maximum JSD achievable given both the number of classes and the residual entropy budget at a given conservation.

Specificity-anchor scoring and trend anomaly

Each position's raw specificity score is the product of between-class divergence and within-class conservation (sdp_score = JSD × conservation), rewarding the conserved-yet-divergent corner. Scores were stabilized by bootstrap resampling of sequences (frequency of appearing in the top-ranked set over resamples; sdp_ness is the bootstrap-normalized anchor strength). Core anchors are the positions meeting all three criteria: upper-right corner membership, bootstrap selection frequency > 80%, and top-ranked anchor weight; these are 3.21, 4.59 and 5.65. The anomaly score, denoted S_sdp, the symbol used for it in the companion preprint P3, is the residual of between-class divergence against the linear conservation–divergence trend (observed minus trend-predicted divergence), isolating positions that diverge more than their conservation predicts.

Phylogenetic gatekeeping

To control for phylogeny, we computed for each position the conditional mutual information I(residue; coupling | subfamily) using GPCRdb subfamily as the conditioning variable, and compared the observed value to a null generated by permuting coupling labels within each subfamily (preserving subfamily composition and finite-sample estimation bias, which cancels on subtraction). Positions were called significant by Benjamini–Hochberg FDR < 0.05; 141 of 212 passed.

Structural and physicochemical analysis

Minimum heavy-atom distances to Gα were computed from the 5-HT2A–mini-Gαq complex (PDB 6WHA), the only structure of the two carrying a bound G protein, and defined each anchor's contact category (<5 Å direct contact, 5–8 Å interface shell, >8 Å buried). The bare active-state 5-HT2A structure (PDB 8ZMG), which has no bound G protein, was used separately for axial/spatial localization of the anchors along the membrane normal (22 of 26 anchors locatable). Per-class modal side-chain volume, hydropathy and charge were computed from the class residue distributions; the between-class range of each property quantifies whether specificity is steric or electrostatic.

Coupling classifier

A random-forest classifier predicted primary coupling (Gs/Gi/o/Gq/11) from residue identity at defined position sets (orthosteric pocket = 14 positions; interface = 53; combinations and anchor subsets as noted), evaluated by leave-one-out cross-validation and reported as overall accuracy, balanced accuracy, and a per-class confusion matrix; significance was assessed by label permutation. The 0.78 balanced accuracy (82% overall) reported for the interface set is against a 0.53 majority baseline.

Coupling entropy and Boltzmann interpretation

Quantitative coupling efficacies (transduction coefficients, log(Emax/EC50)) for 178 receptors were converted to normalized probability distributions over G-protein families; coupling entropy (bits) and the dominance margin ($p_1 - p_2$) summarize each receptor's selectivity. The Boltzmann framing is a phenomenological interpretation that treats $P(G_i) \propto \exp(-\Delta G_i/kT)$

with kT ≈ 0.6 kcal/mol at body temperature; energetic magnitudes quoted are order-of-magnitude readings from the probability ratios, not measured ΔΔG.

Evolutionary analyses

Cross-species drift is the within-receptor Shannon entropy of a position across orthologs. Class-mode turnover counts changes in the class-specific modal residue across four evolutionary layers (primate → mammal → bird/reptile → fish), with coupling labels propagated along orthologs (99% self-consistency). Ancestral states were reconstructed by Fitch parsimony on the 235-sequence tree; the per-position event count is the number of inferred substitutions. All evolutionary results are reported as rank orders of lability, not calibrated times.

Reprogramming test

Starting from 117 Gi/o receptors, positions were mutated toward Gs modal residues in spectrum-recommended order (by anchor rank) and in random order as a control; the coupling classifier's mean P(Gs) was tracked as positions were added. This is an in silico test on sequence fingerprints; "flip" denotes a change in the model's majority-class call.

Software and data

Analyses used Python (NumPy, SciPy, scikit-learn, Biopython) and standard alignment/tree tools. Per-figure source data are provided as Dataset S1 (master position table and per-analysis CSVs); the master table columns are documented in the accompanying README.

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, 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 table 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 P2 and P3, which use the same instrument).

References

All DOIs below, including the bioRxiv preprint, were verified as resolving against CrossRef (registration-agency check, 2026-07); for journal articles the author lists, journal, volume and pages were taken from the resolved CrossRef record. bioRxiv entries are preprints and are flagged as such; cite the peer-reviewed version when available.

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

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA. 2011;108(49):E1293–E1301. doi:10.1073/pnas.1111471108

Teşileanu T, Colwell LJ, Leibler S. Protein sectors: statistical coupling analysis versus conservation. PLoS Comput Biol. 2015;11(2):e1004091. doi:10.1371/journal.pcbi.1004091

Monod J, Wyman J, Changeux JP. On the nature of allosteric transitions: a plausible model. J Mol Biol. 1965;12(1):88–118. doi:10.1016/S0022-2836(65)80285-6

Khersonsky O, Tawfik DS. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu Rev Biochem. 2010;79:471–505. doi:10.1146/annurev-biochem-030409-143718

McBride JM, Eckmann JP, Tlusty T. General theory of specific binding: insights from a genetic-mechano-chemical protein model. Mol Biol Evol. 2022;39(11):msac217. doi:10.1093/molbev/msac217

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–589. doi:10.1038/s41586-021-03819-2

Flock T, Hauser AS, Lund N, Gloriam DE, Balaji S, Babu MM. Selectivity determinants of GPCR–G-protein binding. Nature. 2017;545(7654):317–322. doi:10.1038/nature22070

Flock T, Ravarani CNJ, Sun D, Venkatakrishnan AJ, Kayikci M, Tate CG, et al. Universal allosteric mechanism for Gα activation by GPCRs. Nature. 2015;524(7564):173–179. doi:10.1038/nature14663

Venkatakrishnan AJ, Deupi X, Lebon G, Tate CG, Schertler GF, Babu MM. Molecular signatures of G-protein-coupled receptors. Nature. 2013;494(7436):185–194. doi:10.1038/nature11896

Inoue A, Raimondi F, Kadji FMN, Singh G, Kishi T, Uwamizu A, et al. Illuminating G-protein-coupling selectivity of GPCRs. Cell. 2019;177(7):1933–1947. doi:10.1016/j.cell.2019.04.044

Avet C, Mancini A, Breton B, Le Gouill C, Hauser AS, Normand C, et al. Effector membrane translocation biosensors reveal G protein and β-arrestin coupling profiles of 100 GPCRs. eLife. 2022;11:e74101. doi:10.7554/eLife.74101

Hauser AS, Avet C, Normand C, Mancini A, Inoue A, Bouvier M, et al. Common coupling map advances GPCR-G protein selectivity. eLife. 2022;11:e74107. doi:10.7554/eLife.74107

Pándy-Szekeres G, Caroli J, Mamyrbekov A, Kermani AA, Keserű GM, Kooistra AJ, et al. GproteinDb in 2024: new G protein-GPCR couplings, AlphaFold2-multimer models and interface interactions. Nucleic Acids Res. 2024;52(D1):D466–D475. doi:10.1093/nar/gkad1089

Singh G, Inoue A, Gutkind JS, Russell RB, Raimondi F. PRECOG: PREdicting COupling probabilities of G-protein coupled receptors. Nucleic Acids Res. 2019;47(W1):W395–W401. doi:10.1093/nar/gkz392

Xu P, Huang S, Zhang H, Mao C, Zhou XE, Cheng X, et al. GPCRs steer Gi and Gs selectivity via TM5-TM6 switches as revealed by structures of serotonin receptors. Mol Cell. 2022;82(14):2681–2695. doi:10.1016/j.molcel.2022.05.031

Hilger D, Masureel M, Kobilka BK. Structure and dynamics of GPCR signaling complexes. Nat Struct Mol Biol. 2018;25(1):4–12. doi:10.1038/s41594-017-0011-7

Lane JR, May LT, Parton RG, Sexton PM, Christopoulos A. A kinetic view of GPCR allostery and biased agonism. Nat Chem Biol. 2017;13(9):929–937. doi:10.1038/nchembio.2431

Conflitti P, Lyman E, Sansom MSP, Hildebrand PW, Gutkind JS, Rosenkilde MM, et al. Functional dynamics of G protein-coupled receptors reveal new routes for drug discovery. Nat Rev Drug Discov. 2025;24(4):251–275. doi:10.1038/s41573-024-01083-3

Ballesteros JA, Weinstein H. Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. Methods Neurosci. 1995;25:366–428. doi:10.1016/S1043-9471(05)80049-7

Kooistra AJ, Mordalski S, Pándy-Szekeres G, Esguerra M, Mamyrbekov A, Munk C, et al. GPCRdb in 2021: integrating GPCR sequence, structure and function. Nucleic Acids Res. 2021;49(D1):D335–D343. doi:10.1093/nar/gkaa1080

Fredriksson R, Lagerström MC, Lundin LG, Schiöth HB. The G-protein-coupled receptors in the human genome form five main families. Mol Pharmacol. 2003;63(6):1256–1272. doi:10.1124/mol.63.6.1256

Preprints (not peer-reviewed):

Allostery between distant structural regions dictates selectivity in GPCR:G protein coupling. bioRxiv. 2025. [Preprint — not peer-reviewed] doi:10.1101/2025.08.04.668482

Supporting Information

Supplementary Figures

S1 Fig. Family-slope decomposition. Per-subfamily conservation–divergence slopes, showing the negative conservation–divergence trend is recovered within individual receptor subfamilies and is not an artifact of pooling across the family. (Source figure: family-slope analysis, artifact 14feeb9d-0835-4f79-9b0e-38627ea75931.)

S2 Fig. Per-position cross-species drift map. The full per-position drift profile across the 212 positions (the exploratory version of the axis summarized in Fig 4A), with allosteric

switches and core anchors annotated. (Source: per-position drift, 3f6feea1-c653-4633-af97-53b627209ffa.)

S3 Fig. Four-layer class-mode turnover, full detail. Layer-by-layer (primate → mammal → bird/reptile → fish) class-mode assignments for all positions, underlying the turnover summary in Fig 4B. (Source: layer turnover, 5ffc70b8-a870-4071-8b3b-393f6dbe15a2.)

S4 Fig. Ancestral-trajectory reconstructions. Per-anchor parsimony trajectories on the 235-sequence tree for the core and edge anchors, underlying the event counts in Fig 4C. Presented as qualitative rank-order evidence of evolutionary lability. (Source: ancestral trajectories, 14ebfea9-4bfd-4650-8bd7-15b1cf7bf316.)

S5 Fig. Coupling-probability spectrum, per-receptor detail. The per-receptor coupling probability distributions underlying the entropy spectrum in Fig 5A, ordered by entropy. (Source: coupling-probability spectrum, 1b6b3e6c-64e3-44a4-9c73-3118695cc344.)

S6 Fig. Position 5.66 multi-method portrait. The standalone exploratory portrait of the edge anchor 5.66 (cross-species drift, bootstrap stability, ancestral events, per-class residue composition), summarized in Fig 4D. (Source: 5.66 portrait, dfd6854e-8ef6-4857-9ae6-c8da2cd88f45.)

Supplementary figures are provided as the original exploratory renderings referenced by artifact ID; they can be reformatted to journal style on request. Main-text figures (Fig 1–6) are the publication-grade replots.

Dataset S1. Source data

A single archive containing the master position table and all per-analysis source tables used to generate Fig 1–6. All coordinates use GPCRdb generic numbering (GPCRs) or Pfam match-state numbering (enzymes).

File	Rows	Content	Used in
sdp_master_table_continuous.csv	212	Per-position master table: conservation, entropy, between-class JSD, anchor score/weight, bootstrap stability, sdp-ness, rank, coupling importance, per-class modal residues, anomaly score, conditional-MI columns, phylogenetic-gatekeeping flag	Fig 1, 2
grouping_specificity.csv	209	Between-group JSD under four grouping logics (coupling, ligand, bias, phylogeny) + best grouping	Fig 2
sdp_physicochem.csv	—	Per-position, per-class side-chain composition, mean hydropathy, volume, charge	Fig 3
sdp_gprotein_dist.csv	—	Minimum distance to Gα (5-HT2A–mini-Gαq) and contact category per anchor	Fig 3
coupling_entropy.csv	178	Per-receptor coupling probability distribution, entropy, dominant-pathway probability, dominance margin	Fig 5

per_position_drift.csv	—	Per-position cross-species (ortholog) drift	Fig 4
layer_turnover_4way.csv	—	Four-layer class-mode turnover	Fig 4
asr_events.csv	211	Fitch-parsimony ancestral substitution-event counts per position	Fig 4
crossclass_transfer.csv	25	Class-A vs class-B/C/F specificity divergence for alignable positions	Fig 6
protease_spectrum.csv	208	Serine-protease conservation/JSD/anomaly per Pfam column, with modal residues	Fig 6
kinase_spectrum.csv	—	Protein-kinase conservation/JSD/anomaly per Pfam column, with modal residues	Fig 6
derived_numbers.json	—	Confusion matrix, reprogramming trajectory summary, coupling quantitation	Fig 3, 5
README	—	Column dictionary for the master table and all derived tables	all

Master table artifact: 35ad9064-0572-4768-a42d-ea2d0e32ba9f; README: 1ea1e8e3-55d2-417c-aa78-e2479cef01e0.