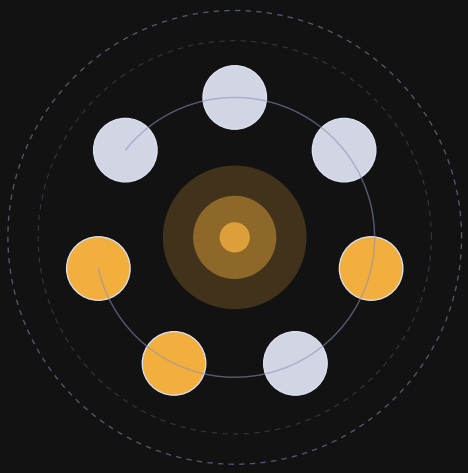
orb::ion



Model Performance Series

Astra AI on GPCRs

How Orbion's six AI models perform across 2,615 G protein-coupled receptors.

AUTHORS

Çağlar Bozkurt[1,*], Aniruddh Goteti[1]

[1]Orbion GmbH, Berlin, Germany

[*]Correspondence: caglar.bozkurt@orbion.life

June 9, 2026 · orbion.life

Executive Summary

G protein-coupled receptors are the largest family of drug targets in the human proteome — roughly a third of marketed small-molecule drugs act on one [11]. Yet GPCRs are also the class where computational characterization fails most often: deep, partly buried orthosteric pockets; conformationally plastic transmembrane bundles; intracellular loops regulated by dense post-translational modification; thermal stability that requires laborious detergent-screen assays to even measure.

Orbion's Astra AI Suite helps GPCR teams **reduce the experimental search space** — prioritizing where to spend reagents and assay time across construct design, thermostability screening, PTM interpretation, and binding-pocket triage. This document is a transparent report on how the suite — six AI models — performs on G protein-coupled receptors. We ran every Astra AI model on **2,615 reviewed GPCRs** from UniProt Swiss-Prot and compared each output against the strongest publicly available experimental reference: Swiss-Prot literature annotation for sequence-level features [4], PDB co-crystal contacts at 4 Å resolution for ligand binding [5], and **154 mutations across 4 receptors** with experimentally measured thermal-shift data for stability prediction [7, 8, 15]. Each section reports the headline performance for one model, its known weaknesses, and where it is recommended for production use.

Headline Numbers Across Six AI Models on GPCRs

- **Residue-Level Transmembrane Topology.** Per-residue prediction agrees with UniProt-annotated transmembrane segments at AUROC 0.97, AUPRC 0.96, F1 0.91. The disorder model reaches AUROC 0.93 on receptors with annotated disordered regions.
- **Post-Translational Modification (PTM) Site Prediction.** F1 up to 0.94 for N-linked glycosylation and disulfide bonds, 0.78 for S-palmitoylation. All 39 modification classes covered; two operating points reported (high-precision and high-recall) for each.
- **Ligand Binding Pocket Prediction.** 64 % recall on PDB-observed ligand identities; 72 % pocket-success rate (the predicted pocket overlaps a known ligand-contact residue set) on the 223 receptors with co-crystal experimental data — *pocket-level triage, not atomic-contact prediction.*
- **Thermostability Prediction (ΔT_m).** On strong-effect destabilizing mutations (experimental $\Delta T_m < -2\,^{\circ}\mathrm{C}$, n = 50): 82 % sign accuracy — usable as a pre-screen to remove risky variants before CPM thermal-shift assays. Across all 154 mutations on four receptors, scored against *independent* experimental data: MAE 2.71 °C, aggregate sign accuracy 69 % — on the same order as the experimental uncertainty of CPM-style measurements [13].
- **Protein Function and Family Classification.** 99 % correctly identified as Receptors (n = 2,615); GO molecular-function term recovered (ancestor-closure) in the top-5 predictions for 99.2 % of receptors, top-1 89.3 %; the enzyme-vs-non-enzyme call has near-perfect calibration.
- **Topology and Membrane Class.** 99.9 % correctly classified as 7-transmembrane on canonical GPCRs (n = 2,536); 100 % correctly classified as multi-pass membrane proteins. (These classification calls are reliable triage on curated sequences rather than the suite's core scientific differentiation — see §6.)

The story below the headlines is uneven — and we say where. Aminergic Class A receptors are the strongest family for binding-pocket prediction and the weakest for phosphorylation site prediction. Glycoprotein-hormone and adhesion receptors, where ligand binding occurs in an extracellular domain rather than the canonical transmembrane pocket, are outside the current scope of the binding model. Below experimental measurement precision ($|\Delta T_m| < 2\,^\circ\mathrm{C}$), thermostability predictions are not actionable signal — not because the model fails, but because the measurement does. We document each of these in §4.

The single most important section of this paper is §3, where we walk every Astra AI model through one receptor — the adenosine A2A receptor (UniProt P29274) — and show what a program team would do with the integrated output in practice. The rest of the paper is the per-model evidence supporting that picture, and a decision framework (§5) that maps five common GPCR program questions to the model combinations that help answer them.

What This Document Is. A performance report on Orbion's Astra AI Suite on the GPCR class. Every number traces to a checked-in source-data artifact and is reproducible from the cited experimental reference. The classification, topology and PTM metrics characterize how the deployed models behave on the GPCR class in production — including proteins drawn from their training corpora; held-out generalization is reported in the per-model preprints. Thermostability (ΔT_m) is the exception, evaluated on independent experimental data (§6).

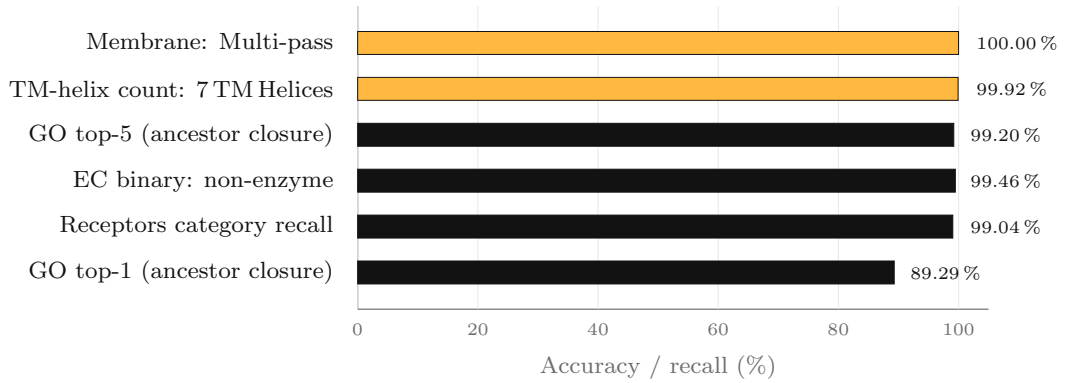


Figure 1: *Headline classification performance on the 2,615-GPCR cohort. The top two bars (amber) are the topology and membrane-class predictions; the four bars below are the function and family classification outputs. Full per-model numbers are reported in §2.1 and §2.2.*

1 Why GPCRs

G protein-coupled receptors form the largest superfamily of integral membrane proteins in the human genome and the most pharmaceutically exploited target class. Roughly a third of FDA-approved small-molecule drugs act through a GPCR [11], and the family continues to expand as recently de-orphanized receptors enter active discovery pipelines. Despite this commercial weight, GPCRs remain one of the harder protein classes to characterize computationally, for reasons that are structural rather than incidental.

Structural Complexity. Each canonical GPCR threads seven transmembrane α-helices through the lipid bilayer, forming an orthosteric pocket that is often deeply buried, partly occluded by extracellular loops, and conformationally plastic between active and inactive states. Allosteric pockets layered onto the transmembrane bundle and at the

lipid interface broaden the druggable surface but complicate prediction. Class B, C, and F receptors deviate further from the canonical Class A scaffold and introduce additional extracellular domains that often participate directly in ligand binding.

Regulatory Complexity. GPCR signaling is shaped post-translationally as much as it is structurally. Receptor desensitization is governed by phosphorylation of the C-terminal tail and the third intracellular loop by GRK and second-messenger kinases; trafficking depends on N-linked glycosylation in the extracellular N-terminus; membrane retention depends on cysteine palmitoylation in helix 8 and the proximal C-tail. Predicting the relevant modification sites from sequence is necessary to interpret biased signaling, agonist-dependent internalization, and isoform-specific behavior in disease tissue.

Experimental Bottlenecks. Thermal-shift assays on detergent-solubilized GPCRs — the standard reagent for characterizing thermostabilization mutants ahead of structural or biophysical work [13] — are laborious, reagent-intensive, and difficult to scale. A mutational landscape of any size (hundreds to thousands of variants per receptor) cannot be enumerated experimentally; some computational prefilter is required.

The Untapped Opportunity. These bottlenecks leave much of the family underexploited. Over a hundred GPCRs remain orphans with no confirmed endogenous ligand; many therapeutically interesting receptors have wild-type sequences too unstable to express, purify, or crystallize, so no structure and no screening campaign ever begins. Each such target that can be characterized and stabilized opens a path to a tractable construct, a structure, and a drug-discovery program where none existed — the difference between an intractable target and a pipeline asset often comes down to a handful of stabilizing mutations and a correct read of its pockets and modification sites. Compressing that read from months of trial-and-error to minutes is where computational prediction earns its place.

Why a Unified AI Approach. The Astra AI Suite is built on a shared protein-language-model foundation — a common sequence representation feeding task-appropriate model architectures for each prediction problem. The suite was validated on this premise: one consistent input pipeline, evaluated transparently per protein family. The remainder of this document reports how the suite performs on the 2,615 reviewed GPCRs in UniProt Swiss-Prot [4], one model at a time, with the experimental reference data and metrics for each.

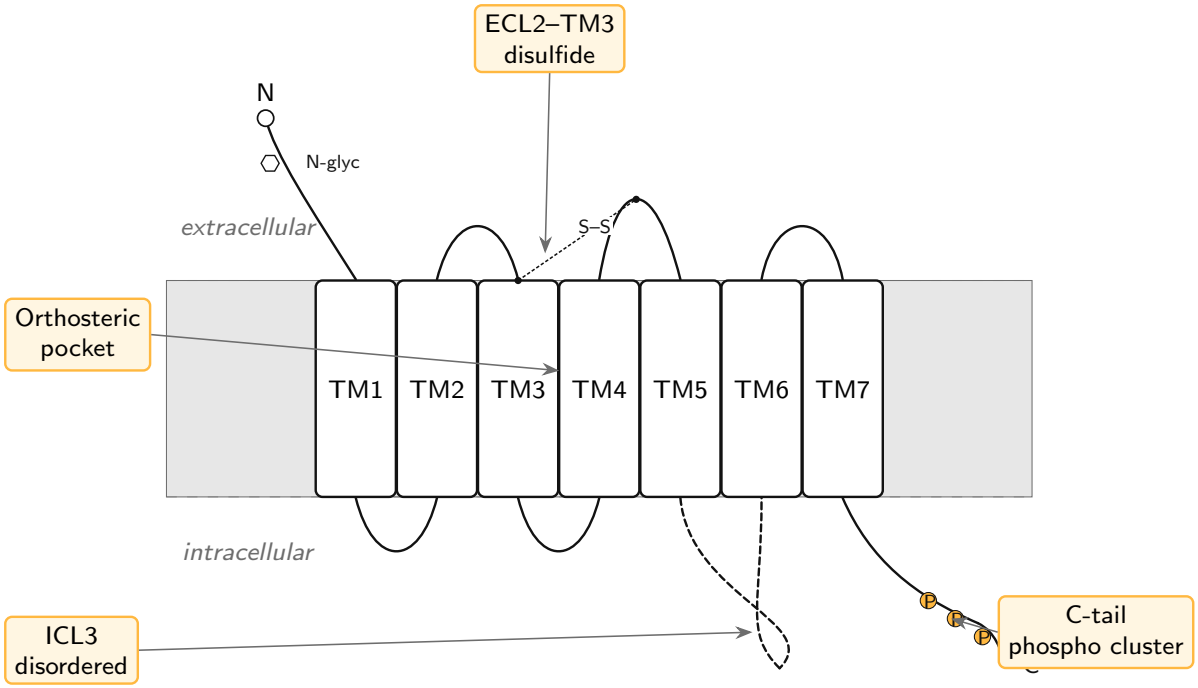


Figure 2: *Canonical Class A GPCR architecture and the prediction tasks the Astra suite addresses. Seven transmembrane α-helices traverse the lipid bilayer, an orthosteric pocket sits within the bundle, the second extracellular loop is anchored to TM3 by a conserved disulfide bond, the third intracellular loop is frequently disordered, the N-terminus carries glycosylation sites, and the C-terminal tail carries the regulatory phosphorylation cluster targeted by GRK and second-messenger kinases.*

2 The Astra AI Suite, Model by Model

The Astra AI Suite is six models that share a common protein-language-model foundation (a sequence representation augmented with structural and physicochemical features), each built on the machine-learning technique best suited to its task — spanning transformer networks, graph neural networks, and regression models. Each output below was evaluated on the 2,615-GPCR cohort against the strongest publicly available experimental reference: Swiss-Prot literature annotation [4], PDB co-crystal heavy-atom contacts at 4 Å resolution [5], Gene Ontology with ancestor-closure semantic matching [9, 10], and experimentally measured thermal shifts for stability. Methodological details, metric definitions, evaluation scope, and per-model cohorts are in §6.

Across the suite, the models are calibrated to be conservative: where signal is weak they tend to omit a prediction rather than fabricate one, so the dominant failure mode is a missed call rather than a false positive — the safer error for wet-lab triage, where a fabricated hit wastes reagents but an omission simply means no guidance. For binding-site prediction this property is reinforced by a seven-gate anti-hallucination audit (§2.5).

2.1 Protein Function and Family Classification

What We Predict. From a protein sequence alone, the protein's broad functional category (Receptors, Transporters, Enzymes, Signaling Proteins), its Enzyme Commission family, its molecular-function Gene Ontology terms, and its pathway memberships. Intended use is rapid sequence triage at the front of a discovery pipeline. Evaluated on all 2,615 reviewed GPCRs in the cohort.

Headline Numbers.

- **Receptors category recall: 99.04 %** (2,590 / 2,615); mean predicted probability when correct: 0.998.
- **Enzyme-vs-non-enzyme accuracy: 99.46 %**; **Brier score: 0.005** (a measure of calibration; 0 is perfect, 0.25 is uninformed).
- **GO molecular-function top-5 (ancestor closure): 99.20 %**; top-1: 89.29 %.

Strong and Weak. Performance is uniform across GPCR sub-families: Class A aminergic, peptide, chemokine, opsin, olfactory; Class B1 secretin; Class B2 adhesion; Class C glutamate-like; Class F frizzled. Recall on Receptors falls between 99 % and 100 % in every family stratified. One outlier: a single GPCR was confidently classified as a hydrolase at P = 0.77.

Use as: production triage — reliable enough to gate a discovery pipeline and assign broad functional context.

2.2 Topology and Membrane Class

What We Predict. A set of auxiliary structural and biochemical classifications: cofactor requirements, taxonomic domain and host, membrane topology (soluble / single-pass / multi-pass), subcellular localization, transmembrane-helix count bucket, and quaternary structure. The transmembrane-helix-count and membrane-class outputs were scored on the 2,536 canonical multi-pass GPCRs with seven UniProt-annotated transmembrane helices.

Headline Numbers.

- **Transmembrane-helix count: 99.92 % correctly classified as "7 TM Helices"** (2,534 / 2,536).
- **Membrane class: 100 % correctly classified as "Multi-pass"** (2,536 / 2,536).

Strong and Weak. The two errors on transmembrane-helix count are atypical Class A sequences (likely partial-length or splice-variant records). Uniform across families. The remaining classifiers (cofactor, host, subcellular, quaternary) require external experimental reference and will be reported in a supplement.

Use as: production triage — structural-class confirmation before modeling or pipeline routing.

2.3 Residue-Level Topology and Disorder

What We Predict. For each residue in a protein sequence, the probability of lipid-bilayer embedding, intracellular orientation, extracellular orientation, intrinsic disorder, and amyloidogenicity. Evaluated on 2,600 GPCRs with UniProt transmembrane annotations and 803 GPCRs with UniProt disorder annotations.

Headline Numbers.

- Transmembrane prediction: per-protein **mean AUROC 0.969, AUPRC 0.958**, F1 at threshold-optimal cutoff 0.909, Brier 0.075.
- Disorder prediction (n = 803): mean AUROC 0.930, AUPRC 0.611.

Strong and Weak. Transmembrane prediction is excellent and uniform across canonical receptor classes (per-family mean AUROC 0.97–0.99 across Class A peptide / aminergic / chemokine / opsin, Class B1 / B2, Class C, Class F). Modest degradation on bitter taste (TAS2R, mean AUROC 0.910) and vomeronasal (mean AUROC 0.870) sub-families, both with substantially sparser Swiss-Prot annotation.

Use as: production triage for transmembrane topology; hypothesis generation for the disorder prediction.

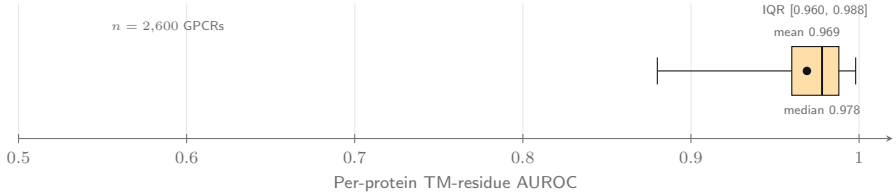


Figure 3: *Distribution of per-protein transmembrane-residue AUROC across the 2,600 GPCRs with UniProt transmembrane annotations.*

	n	Receptors recall	TM-residue AUROC	Phospho AUROC	Phospho R@n
Class A — Opsins	134	**1.000**	**0.990**	**0.999**	**0.963**
Class A — Chemokine	133	0.977	0.977	0.956	0.692
Class A — Peptide	369	0.995	0.971	0.899	0.501
Class A — Aminergic	258	0.992	0.987	0.775	0.088
Class A — Orphan (GPRn)	180	1.000	0.964	0.890	0.406
Olfactory (OR*)	524	1.000	0.957	*n/a*	*n/a*
Taste 2 (TAS2R)	153	1.000	0.910	*n/a*	*n/a*
Class B1 — Secretin	62	1.000	0.982	0.858	0.204
Class B2 — Adhesion (ADGR)	55	1.000	0.991	0.907	0.300
Class C — Glutamate-like	57	1.000	0.991	0.901	0.338
Class F — Frizzled / SMO	43	1.000	0.992	0.805	0.611

Cell shade ∝ value (white = low, amber = high). *n/a* = no Swiss-Prot phosphorylation annotations for this family.
Phospho R@n = recall when top-n predicted residues are chosen, n = real annotated sites per protein.

Figure 4: *Performance across the GPCR receptor super-family, broken down by sub-family. Cells shaded by value (white = low, amber = high). Receptor classification is uniformly excellent and transmembrane prediction degrades only modestly on bitter taste receptors. Phosphorylation is shown as the representative post-translational modification because it has the broadest per-family ground-truth coverage and by far the largest family-dependent variation (near-perfect on opsins, weak on the heavily-phosphorylated aminergic receptors); the strongest modification classes — disulfide bonds and N-linked glycosylation — perform uniformly well across families and are reported per-class in Figure 5.*

2.4 Post-Translational Modification (PTM) Site Prediction

What We Predict. For each residue, the probability that the residue carries any of 39 post-translational modifications. Two operating points are emitted in parallel: a high-precision call calibrated for confident wet-lab handoff, and a high-recall call calibrated for

hypothesis generation [1]. The dual-output design responds to the incomplete-annotation regime characteristic of PTM literature. Evaluated on the subset of the cohort with ≥ 1 Swiss-Prot-curated positive per modification class.

Headline Numbers. Figure 5 reports F1 at the threshold-optimal operating point per modification class. The high-recall operating point dominates the high-precision operating point on six of seven classes with sufficient ground truth.

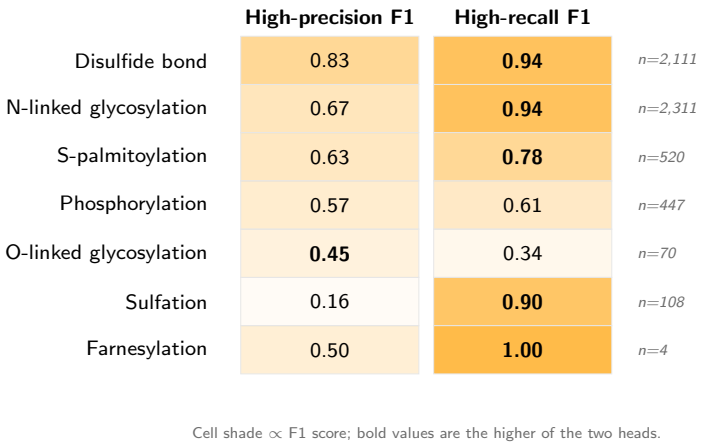


Figure 5: *PTM site prediction F1 at the threshold-optimal operating point on GPCR Swiss-Prot experimental reference, per modification class. O-linked glycosylation is the lone inversion.*

Strong and Weak. Strongest on opsins for phosphorylation: the tightly clustered C-terminal phospho-cluster is well-captured (recall at $k = \mathrm{n}_{\mathrm{positive}}$ of 0.96). Weakest on Class A aminergic receptors for phosphorylation (adrenergic, dopamine, serotonin, histamine, muscarinic — carrying long, heavily-phosphorylated C-terminal tails regulated by GRK1–6, PKA, PKC, with 10+ annotated sites per protein). The conservative output budget limits candidates per protein, so k is exhausted before all real sites are recovered — consistent with the suite-wide tendency to miss rather than misflag. Disulfide bond and N-linked glycosylation perform broadly strongly across every GPCR sub-family.

Use as: production triage at the high-precision operating point (confident wet-lab handoff — mass-spectrometry targeting, mutagenesis prioritization); hypothesis generation at the high-recall operating point.

2.5 Ligand Binding Pocket Prediction

What We Predict. Given a target protein sequence, a ranked panel of plausible ligand candidates with each ligand's predicted binding-residue set [3]. The candidate panel is drawn from established ligand resources (the RCSB Chemical Component Dictionary, AlphaFill homology placements, and a curated catalog), so the model's role is *retrieval, ranking, and pocket localization of known ligands* — not de-novo proposal of novel chemistry. A seven-gate anti-hallucination audit verifies position validity, confidence-score bounds, ligand-ID realism, reproducibility, and biological plausibility on every production batch. Evaluated on the 223 GPCRs with both predictions and PDB co-crystal contacts at 4 Å resolution.

Headline Numbers. *Read the scope before the numbers:* these metrics measure known-ligand retrieval and pocket localization scored against PDB co-crystal contacts — not atomic-level contact prediction. The model deliberately proposes broader candidate pockets (median 12 ligands per protein) than the strict 4 Å PDB contact sets it is scored against (median 3 per protein), so the residue-level metric below is a conservative lower

bound on pocket utility rather than a contact-prediction score.

- **Ligand-identity recall:** **64 %** — of PDB-observed ligand codes per protein, the model surfaces 64 % of them in its candidate panel.
- **Pocket-success rate:** **72 %** — 72 % of model predictions overlap at least one PDB-observed ligand-contact residue set.
- **Residue F1 on ligand-ID-matched predictions: 0.22** — to be read against the scope note above (broad pocket localization vs. strict 4 Å atomic contacts), not as an atomic-contact-prediction score.

Strong and Weak. Strongest on Class A aminergic and well-studied small-molecule-targeted receptors: HTR1F serotonin (F1 = 0.50 on 2 of 2 matched ligands), HCAR2 (0.47 on 7 of 14), TAAR1 (0.43 on 12 of 14), DRD3 (0.42), ADORA3 (0.39). The aminergic family is the strongest for binding prediction, contrasting with the same family's weakness on PTM phosphorylation. On glycoprotein-hormone and adhesion receptors (TSHR, LGR4, ADGRB3, GPR158), where ligand binding occurs in a large extracellular domain rather than the transmembrane pocket, the model does not extend a confident pocket into that domain — a coverage gap rather than a wrong call (see §4). The residue F1 reflects this scope: the suite is built for pocket localization and ligand hypothesis generation, not atomic-level contact prediction.

Use as: pocket localization and ligand-panel triage on canonical transmembrane-pocket receptors. Extracellular-domain receptor classes (glycoprotein-hormone, adhesion, LGR) are out of the current model's scope.

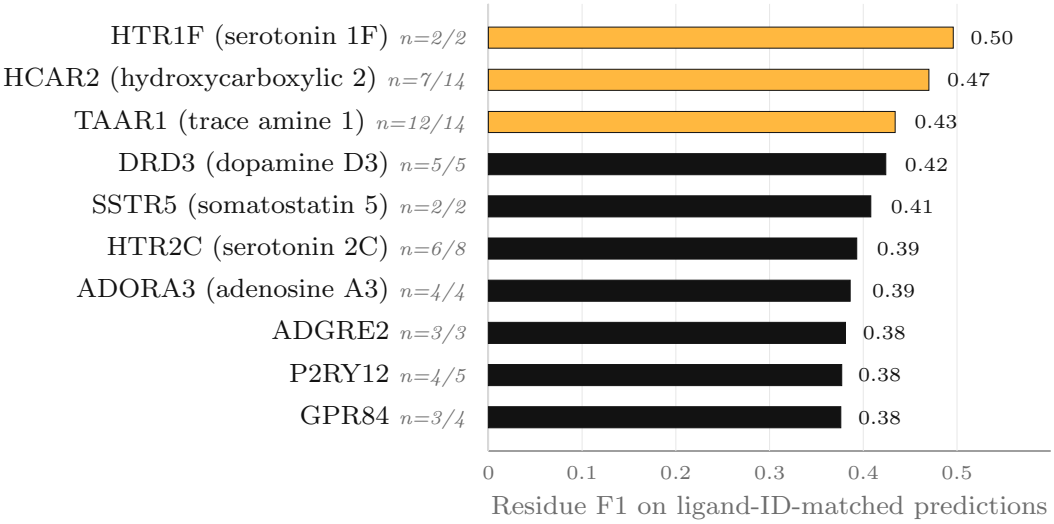


Figure 6: *Top 10 GPCRs by binding-residue F1 on ligand-identity-matched predictions — a pocket-level triage metric, not atomic-contact prediction. The top three (amber) clear F1 ≥ 0.43.*

2.6 Thermostability Prediction (ΔT_m)

What We Predict. For a target protein and a point mutation, the change in melting temperature (ΔT_m) of the mutant relative to wild-type, in degrees Celsius, together with a confidence interval. Evaluated on 154 mutations across four publicly-validated GPCRs (NTSR1 n = 37, A2A n = 26, β_1-AR n = 61, 5-HT2C n = 30), using experimentally measured thermal-shift values as ground truth. Unlike the annotation-recall metrics

elsewhere in this report, this is an *independent* experimental evaluation: the measured ΔT_m values are not part of the sequence models' training corpus, so the numbers below reflect genuine predictive generalization rather than within-distribution recall (see §6).

Headline Numbers.

- **Sign accuracy on strong-effect destabilizing mutations (experimental $\Delta T_m < -2\,$°C, n = 50): 82 %** — the production operating point, usable to filter risky variants before a thermal-shift assay.
- Across all 154 mutations (all four receptors): **MAE 2.71 °C**, Spearman $\rho = 0.47$, aggregate sign accuracy 69 % (stabilizing-vs-destabilizing call).

Note on the Experimental Measurement Floor. The aggregate MAE (2.71 °C) is on the same order as the experimental uncertainty and inter-assay variability of CPM-style thermal-shift measurements, whose between-replicate standard deviation alone is in the ±1.5 °C range [13]. In the small-effect band ($|\Delta T_m| < 2\,$°C, either direction), the prediction is compared against a measurement whose own resolution is comparable, so per-mutation errors there are dominated by experimental noise rather than model error. Performance on strong-effect destabilizing mutations (experimental $\Delta T_m < -2\,$°C) — the practically relevant operating point for filtering risky variants — is higher (82 % sign accuracy) than the aggregate metrics suggest.

Strong and Weak. Strongest on NTSR1, where the mutation set is compact and biologically informative (37 mutations covering the C-terminal thermostabilization positions in literature) — β_1-AR has the larger evaluation set (n = 61). Conservative on extreme stabilizers: the model compresses extreme effects toward the mean (a real +5 °C stabilizer may be predicted at +2 °C; the sign is correct but the magnitude is biased low). On strong-effect destabilizing mutations (experimental $\Delta T_m < -2\,$°C, n = 50), sign accuracy reaches 82 % — useful as a pre-screen to filter the majority of risky variants before a thermal-shift assay. (This figure is specific to destabilizers; the mean-regression on extreme *stabilizers* above means the strong-effect performance is not symmetric.)

Use as: a *ranker* to prioritize candidate mutations by predicted ΔT_m (Spearman $\rho = 0.47$ supports ordering a panel for synthesis); and a production *pre-screen* for strong destabilizers (experimental $\Delta T_m < -2\,$°C, 82 % sign accuracy) to filter risky variants before a thermal-shift assay. Below actionable resolution for small effects ($|\Delta T_m| < 2\,$°C).

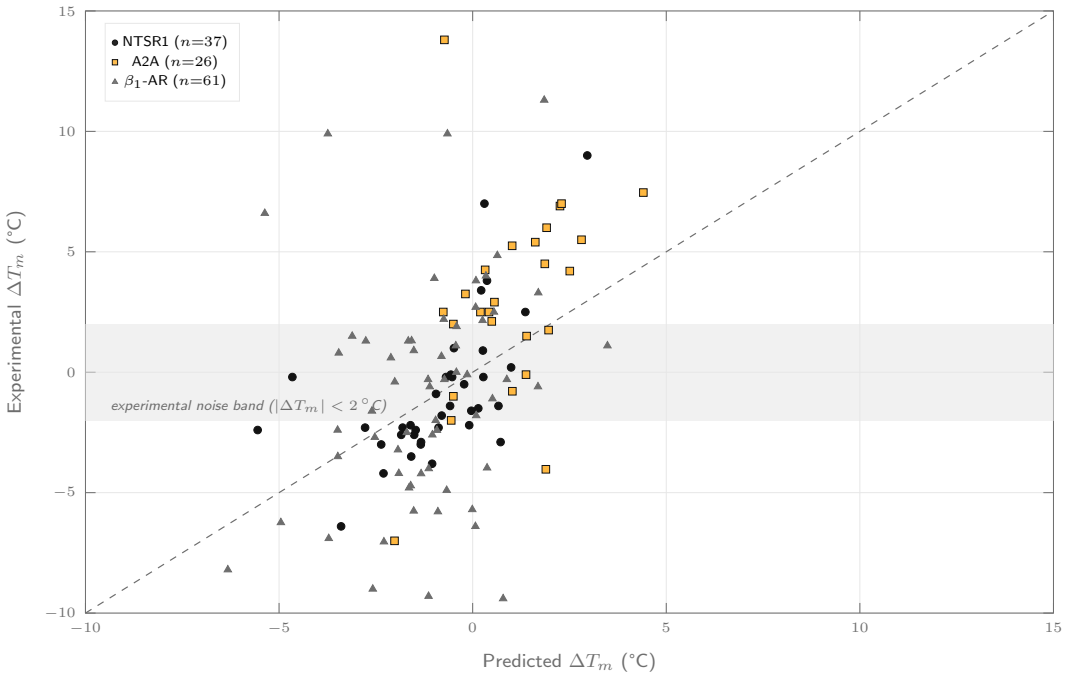


Figure 7: *Predicted vs. experimentally measured ΔT_m. The scatter shows NTSR1, A2A, and β_1-AR (124 mutations); 5-HT2C (n=30) is included in the aggregate statistics above but not plotted individually. The shaded band marks the experimental measurement floor ($|\Delta T_m| < 2\,^\circ C$); the dashed diagonal is the line of perfect prediction.*

3 Worked Example: Adenosine A2A Across the Full Suite

The headline numbers in §2 aggregate across thousands of receptors. In practice, a discovery team works one receptor at a time. This section walks every Astra AI model through a single well-characterized target — the adenosine A2A receptor (UniProt P29274, 412 residues) — and shows what the integrated output looks like at the level of a target a program team is actually working on. A2A has 27 distinct co-crystallized ligands in the PDB experimental reference (§2.5), is the clinical target of istradefylline (Parkinson's disease), and is an active target for cancer immunotherapy — dense crystallographic reference data plus active clinical relevance.

Figure 8 summarizes every per-model output on A2A in a single structural view. The function-and-family classification returns **Receptors at P = 0.998** and **Non-enzyme at P = 1.000**; the GO molecular-function call is an exact match against GO:0004930 G protein-coupled receptor activity. The topology classification returns **7 transmembrane helices at P = 0.998** and **multi-pass membrane at P = 0.999**. The residue-level topology prediction scores above 0.5 in seven contiguous spans corresponding to the UniProt-annotated transmembrane segments, with disorder probabilities elevating in the predicted intracellular loop 3 and proximal C-terminus.

The PTM site prediction returns a compact, biologically defensible set: **five phosphorylation sites at positions 212, 318, 319, 334, and 386** (a cluster in the proximal and distal C-terminal tail consistent with GRK and PKC sites characterized in A2A desensitization literature); **two disulfide-bond positions at 76 and 165** forming the canonical ECL2–TM3 disulfide that stabilizes Class A receptor architecture; and **one N-linked glycosylation site at position 153** in extracellular loop 2. The high-recall output broadens this to twenty phosphorylation candidates — additional positions for

biased-signaling research and site-directed mutagenesis.

The binding-pocket prediction returns a panel of twelve candidate ligands for A2A, including **ADENOSINE** (the endogenous agonist, recovered as a top prediction), **ZMA** (ZM241385, the gold-standard A2A antagonist used in PDB structures 3EML, 3PWH, 3VG9), and **LJX** (a recently disclosed ligand observed in the 2024 structure 8CU6). The predicted binding-residue sets correctly localize the orthosteric pocket to TM3, TM5, and TM6 of the transmembrane bundle.

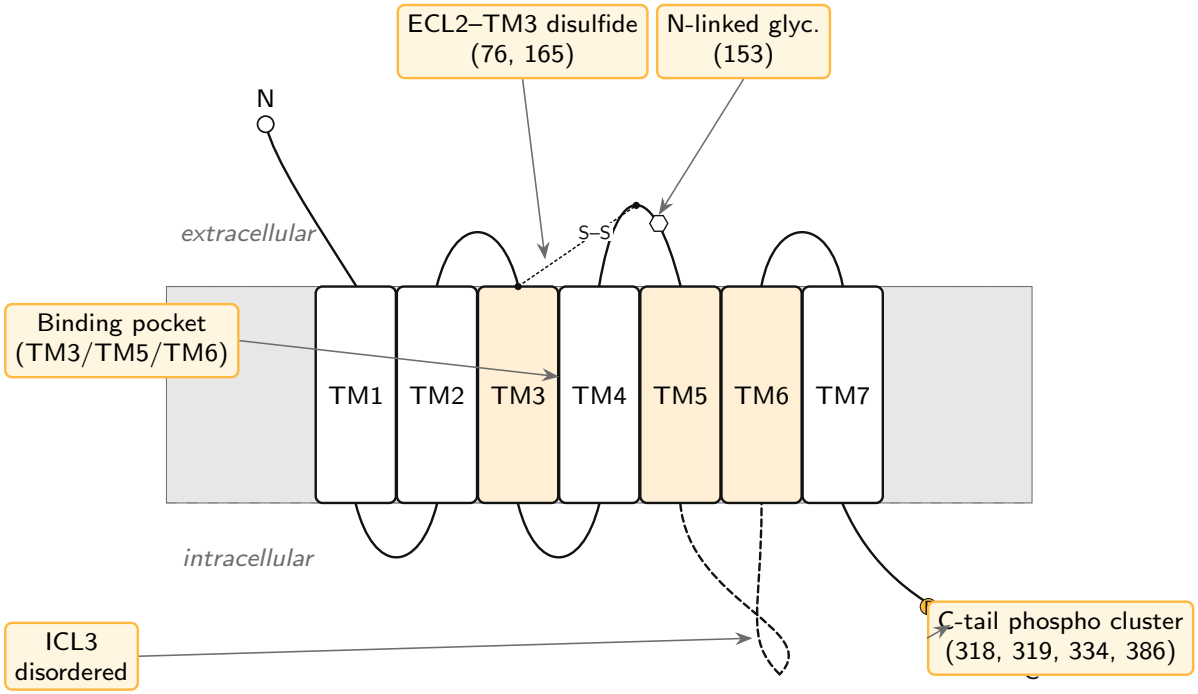


Figure 8: *Integrated view of all model predictions on adenosine A2A (UniProt P29274, 412 residues). Seven transmembrane segments align with the canonical Class A architecture; TM3/TM5/TM6 (amber-tinted) form the predicted orthosteric pocket; the canonical ECL2–TM3 disulfide between positions 76 and 165 is identified with high confidence; one N-linked glycosylation site is predicted at position 153 in ECL2; five phosphorylation candidates cluster in the C-terminal tail; ICL3 is predicted as disordered (dashed). All predictions produced from sequence alone in a single inference pass.*

On the 26 A2A mutations with reported experimental ΔT_m, the model calls the **direction of effect correctly for 76 %** of mutations and is most accurate precisely where construct-design decisions are made — the small-to-moderate effects (V282A: predicted +1.39 °C vs experimental +1.50 °C; S90A: +1.96 vs +1.75 °C). Ranked by predicted ΔT_m, it surfaces the stabilizing candidates a team would prioritize for synthesis. As across the suite, it is conservative on extreme magnitudes: a few large outliers are compressed toward the mean (the +13.8 °C thermostabilizer L48A is the clearest case), so candidate stabilizers flagged by the model should have their magnitudes confirmed experimentally rather than read off the prediction.

What a Program Team Would Do With This Output. A team working A2A from sequence alone, with no prior structural model in hand, can move directly from this output to a defensible construct design and a wet-lab mutation list:

- Preserve the canonical ECL2–TM3 disulfide during any cysteine engineering.
- Retain the N-glycosylation site at position 153 for expression yield.

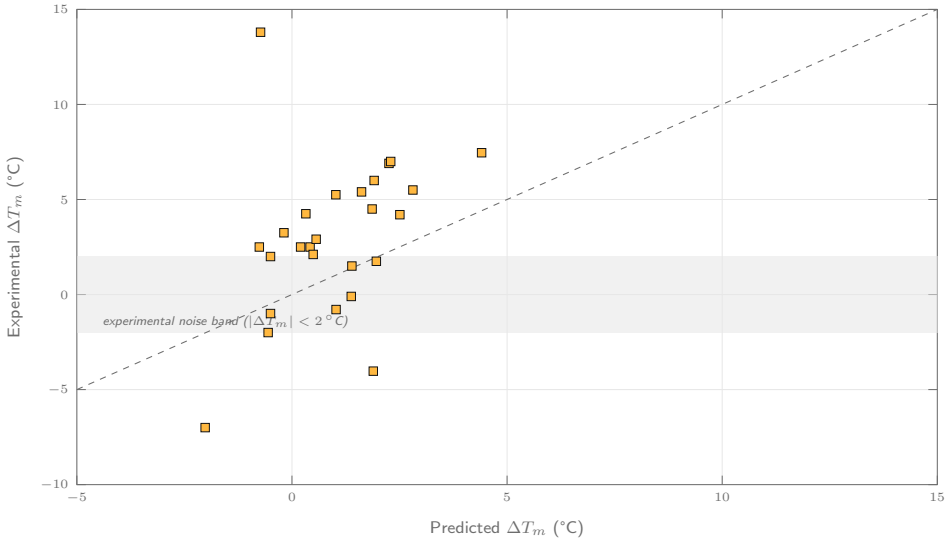


Figure 9: *Predicted vs. experimentally measured ΔT_m on the 26 A2A mutations. Predictions track experiment well through the small-to-moderate range; the dashed diagonal is the line of perfect prediction and the shaded band marks the experimental measurement floor ($|\Delta T_m| < 2\,°C$).*

- Mutate the five high-confidence C-tail phosphorylation sites (S → A) for a desensitization-resistant construct.
- Filter destabilizing variants with the strong-effect ΔT_m predictions (82 % sign accuracy on strong destabilizers across the GPCR set, §2.6) before committing to a CPM screen; confirm magnitude calls on candidate stabilizers experimentally.
- Seed a virtual screening campaign or de-novo binder design from the twelve binding-pocket candidates (including ADENOSINE, ZMA, LJX), focusing mutagenesis on the predicted TM3/TM5/TM6 pocket residues.

The output yields a ∼20–40 variant panel ready for wet-lab follow-up.

4 Limits and Operating Envelope

Two operating boundaries are worth making explicit. Each is a scope statement, not a recovery roadmap.

Binding-Pocket Prediction Scope — Canonical Transmembrane Pockets Only. The 64 % ligand-identity recall and 72 % pocket-success rate in §2.5 are aggregate metrics on the canonical orthosteric and allosteric pocket in the transmembrane bundle. On receptor classes whose primary ligand-binding site is in a large extracellular domain — glycoprotein-hormone receptors (TSHR, LHCGR, FSHR), adhesion-class GPCRs (the ADGR family), and the leucine-rich-repeat-containing receptors (LGR4–6) — the model correctly identifies the ligand from chemistry alone but assigns its residue contacts to the transmembrane pocket and misses the extracellular binding site. These receptors should be excluded from binding-pocket triage in the current version.

Thermostability Prediction Operating Envelope — The Experimental Measurement Floor. The aggregate MAE of 2.71 °C reported in §2.6 is on the same order as the experimental uncertainty and inter-assay variability of CPM-style thermal-shift measurements. Predictions in the small-effect band ($|\Delta T_m| < 2\,°C$, either direction) are being

compared against an experimental measurement whose own resolution is comparable. This is a property of the evaluation, not the model. The strong-effect destabilizing operating point (experimental $\Delta T_m < -2\,^\circ\mathrm{C}$), which is the wet-lab-relevant decision boundary for filtering risky variants, achieves 82 % sign accuracy and is the metric to cite for production use. (Performance is not symmetric: extreme *stabilizers* are subject to mean-regression, as the A2A L48A example in §3 shows.)

5 Decision Framework: Which Models for Which Question

The Astra AI Suite is most useful when used in combination. The table below maps five common GPCR program questions to the model combinations that address them, the operating points to use, and the expected output for a discovery team. Numbers in parentheses reference the per-model evidence in §2.

Table 1: *Decision framework mapping common GPCR program questions to combinations of Orbion's Astra AI Suite. The integrated suite is most valuable when used as a workflow; single-prediction-area use is appropriate for triage but generally leaves information on the table.*

Program Question	Recommended Workflow Across Orbion's Astra AI Suite
1. Thermostabilization Mutation Panel Design	The **thermostability prediction** ranks candidate point mutations by predicted ΔT_m; filter to the strong-effect destabilizing operating point (experimental $\Delta T_m < -2\,^\circ\mathrm{C}$, 82 % sign accuracy). Cross-check the resulting mutation positions against the **residue-level topology prediction** to confirm they fall in the transmembrane bundle rather than disordered loops, and against **PTM site prediction** to avoid introducing mutations at conserved post-translational sites. A 20–40 variant panel for CPM-assay validation is the typical output.
2. Pocket Identification on Novel Targets	The **ligand binding pocket prediction** returns a ranked candidate-ligand panel with predicted binding-residue sets (72 % pocket-success rate on canonical transmembrane-pocket receptors). Cross-check the predicted pocket residues against the **residue-level topology prediction** to distinguish orthosteric (transmembrane) from interfacial pockets. For receptor classes with extracellular-domain binding (glycoprotein-hormone, adhesion), supplement with structural inspection (see §4).
3. Triaging De-Novo Binder Designs	For each candidate binder construct, run the **function and family classification** to confirm the design resembles a receptor-binding entity in functional space; the **residue-level topology prediction** to verify the design has no spurious transmembrane segments; **PTM site prediction** to flag potential N-glycosylation sites that may interfere with the binding interface; and the **thermostability prediction** to estimate the thermal-stability impact of the variable residues. This combination is the integration point between Orbion's Astra AI Suite and antibody / mini-protein / cyclic-peptide design pipelines.

Program Question	Recommended Workflow Across Orbion's Astra AI Suite
4. Loss-of-Function Variant Filtering	The **thermostability prediction** estimates the stability impact of each variant; combine with **PTM site prediction** to flag variants that disrupt canonical modification sites (*e.g.* Ser/Thr-to-Ala at a phosphorylation position) and the **residue-level topology prediction** to identify variants in functionally critical transmembrane positions. ClinVar or gnomAD missense variants can be re-ranked by a combined Astra-derived deleteriousness score for follow-up.
5. Mapping the PTM Landscape for Biased Signaling	The **PTM site prediction** (high-recall operating point) returns the broader candidate-site set across all 39 modification classes; intersect with the **residue-level disorder prediction** to identify modification sites in disordered intracellular loops (the GRK / arrestin substrate regions); intersect with the literature-curated set in Swiss-Prot to surface novel candidate sites not in current annotation. Output is a hypothesis-grade modification map suitable for mass-spectrometry follow-up.

The recurring pattern across all five workflows is the same: a primary model produces a candidate set, and one or two complementary models filter or contextualize the candidates against other protein properties. This is the integration story — Orbion's Astra AI Suite is built so that one sequence-input call returns enough complementary outputs to make multi-criteria decisions in a single pass.

About Orbion

Orbion is an AI-powered protein engineering platform that compresses the workflow from sequence to experiment-ready protocol. The Astra AI Suite described in this document is the prediction layer of the platform — functional annotation, post-translational modification site prediction, ligand binding-site identification, residue-level topology and disorder, and thermostability prediction across ΔT_m and $\Delta\Delta G$. Sitting alongside it on the same platform are AlphaFold2 / AlphaFold-Multimer structure prediction, a construct engineering workspace with composite scoring against an organization's expression vector library, automated experimental protocol generation across expression / purification / crystallization / cryo-EM / stabilization, and a mutation engine for thermostabilization campaigns. Orbion was founded in 2024, is headquartered in Berlin, and works with contract research organizations, pharmaceutical and biotech teams, and academic groups. Free access is available to academic users with a partnership agreement.

Platform. app.orbion.life **Web.** orbion.life **Contact.** `contact@orbion.life` **Academic access.** orbion.life/researcher-program

6 Methods

Cohort. 2,615 reviewed GPCRs from UniProt Swiss-Prot [4] by keyword KW-0297, all organisms, de-duplicated against sequence-integrity checks. Per-area evaluation subsets are stated inline in each §2 subsection.

Experimental Reference. UniProt feature table for sequence-level features (transmembrane segments, modified residues, glycosylation sites, disulfide bonds, lipidation, regions). Gene Ontology molecular-function annotations expanded to the is_a / part_of ancestor closure via QuickGO [9, 10]. PDB ligand-contact reference (§2.5): heavy-atom contacts at 4.0 Å between each ligand and protein residue, aggregated across all deposited co-crystal structures per UniProt accession [5]. ΔT_m reference (§2.6): public literature sources including the Heptares thermostabilization series and reported deep mutational scans.

Metrics. AUROC, AUPRC [14], F1 at threshold-optimal cutoff, F1 at 0.5, recall at $k = \mathrm{n}_{\mathrm{positives}}$, Brier score, MAE in degrees Celsius, Pearson and Spearman correlations, sign accuracy (computed on the subset of mutations with $|\Delta T_m| > 0.5\,^{\circ}\mathrm{C}$). Per-protein metrics aggregated by mean; distributions reported where shape is informative.

Evaluation Scope. The classification, topology, and PTM metrics in §2.1–§2.4 measure how the deployed Astra models perform on the GPCR class as they are run in production. The GPCR cohort defined above (reviewed Swiss-Prot, all organisms) overlaps the corpora these models were trained on; the figures therefore characterize within-distribution behavior on real targets rather than held-out generalization. Held-out test-split performance for each model is reported in its respective preprint [1–3]. The contribution of this report is the per-family GPCR performance breakdown and the integrated single-receptor workflow (§3). The thermostability evaluation (§2.6) is the exception: it scores predictions against an independent experimental thermal-shift dataset that is not part of the sequence models' training corpus, and so reflects genuine predictive generalization.

Reproducibility. Every aggregate number traces to a checked-in source-data artifact in Orbion's evaluation repository, versioned alongside the published methods [1–3].

References

[1] Bozkurt, Ç., Vasilyeva, A., Goteti, A. AstraPTM2: A Context-Aware Transformer for Broad-Spectrum PTM Prediction. *bioRxiv* 2025.10.03.680341 (2025). doi:10.1101/2025.10.03.680341.

[2] Bozkurt, Ç., Vasilyeva, A., Goteti, A. AstraROLE2 & AstraSUIT2: Multi-Task Annotation Models for Functional Profiling of Proteins. *bioRxiv* 2025.06.21.660734 (2025). doi:10.1101/2025.06.21.660734.

[3] Goteti, A., Vasilyeva, A., Bozkurt, Ç. AstraBIND: Graph Attention Network for Predicting Ligand Binding Sites. *bioRxiv* 2025.11.10.687555 (2025). doi:10.1101/2025.11.10.687555.

[4] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, **51**(D1):D523–D531 (2023). doi:10.1093/nar/gkac1052.

[5] Burley, S. K. *et al.* RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research*, **47**(D1):D464–D474 (2019). doi:10.1093/nar/gky1004.

[6] Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, **50**(D1):D439–D444 (2022). doi:10.1093/nar/gkab1061.

[7] Magnani, F., Shibata, Y., Serrano-Vega, M. J., Tate, C. G. Co-evolving stability and conformational homogeneity of the human adenosine A2a receptor. *Proceedings of the National Academy of Sciences*, **105**(31):10744–10749 (2008). doi:10.1073/pnas.0804396105.

[8] Warne, T. *et al.* Structure of a β_1-adrenergic G-protein-coupled receptor. *Nature*, **454**:486–491 (2008). doi:10.1038/nature07101.

[9] Binns, D. *et al.* QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics*, **25**(22):3045–3046 (2009). doi:10.1093/bioinformatics/btp536.

[10] Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**:25–29 (2000). doi:10.1038/75556.

[11] Sriram, K., Insel, P. A. G Protein-Coupled Receptors as Targets for Approved Drugs: How Many Targets and How Many Drugs? *Molecular Pharmacology*, **93**(4):251–258 (2018). doi:10.1124/mol.117.111062.

[12] Pándy-Szekeres, G. *et al.* GPCRdb in 2023: state-specific structure models using AlphaFold2 and new ligand resources. *Nucleic Acids Research*, **51**(D1):D395–D402 (2023). doi:10.1093/nar/gkac1013.

[13] Alexandrov, A. I. *et al.* Microscale fluorescent thermal stability assay for membrane proteins. *Structure*, **16**(3):351–359 (2008). doi:10.1016/j.str.2008.02.004.

[14] Davis, J., Goadrich, M. The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233–240 (2006). doi:10.1145/1143844.1143874.

[15] Esposito, D. *et al.* MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biology*, **20**:223 (2019). doi:10.1186/s13059-019-1845-6.

[16] Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, **46**(D1):D1062–D1067 (2018). doi:10.1093/nar/gkx1153.