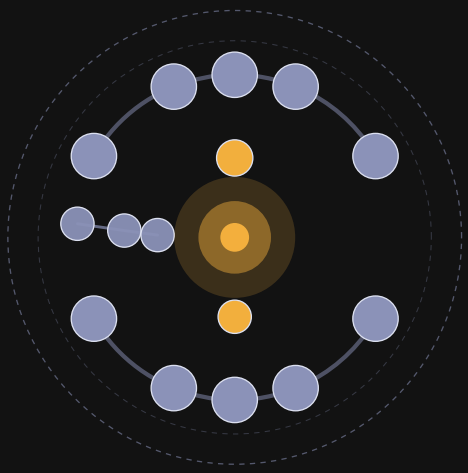
orb::ion



Model Performance Series

# Astra AI on Proteases

How Orbion's five AI models perform across 545 reviewed human proteases.


AUTHORS

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

[1]Orbion GmbH, Berlin, Germany

[*]Correspondence: aniruddh.goteti@orbion.life



June 28, 2026 · orbion.life

## Executive Summary

Proteases are one of the most clinically validated enzyme classes in medicine — from the antihypertensives targeting ACE and neprilysin to the antiviral protease inhibitors — yet the hard targets still fail early on construct, topology, and selectivity risk. This volume is written for protease program leads, structural biology teams, and discovery decision-makers who need to move a difficult target from sequence to a defensible next experiment. Astra does not replace the lab. It narrows the experimental search space before teams commit protein-production, co-crystal, screening, and selectivity budget.

This document reports how Orbion's Astra AI Suite performs across five prediction areas on the human protease class (Figure 1). We ran every Astra model on **545 eligible reviewed human proteases** from UniProt Swiss-Prot [4] and compared each output against the strongest publicly available experimental reference: Swiss-Prot literature annotation for sequence-level features [4], and PDB co-crystal contacts at 4.0 Å resolution for ligand binding [5].

**Headline Numbers Across Five Prediction Areas on Human Proteases**

- **Function and Protease Family.** 91.6 % of the 545 eligible reviewed human proteases are recognised as hydrolases (EC 3), and 90.1 % carry protease/peptidase GO terms.
- **Topology, Cofactor, and Membrane Class.** Membrane-class routing matches UniProt on 81.7 % of the cohort (445/545), with zinc cofactor called on 124 targets.
- **Residue-Level Topology.** Per-residue transmembrane agreement reaches AUROC 0.951 (median 0.995) across 109 annotated membrane proteases.
- **PTM Site Prediction.** F1 0.924 for N-linked glycosylation and 0.829 for disulfide bonds — the construct-design signals that matter most for protease production.
- **Ligand Binding Panel.** 99/242 scored proteins reach ligand-set F1 $\geq 0.5$ on the PDB-contact subset; exact top-1 crystallographic-ID recall (per ground-truth ligand pair) is 0.053, reported separately as the strictest governance check, not the product promise.

**The decision value.** From one sequence, Astra returns the evidence a protease team needs for a first decision review: family and chemistry, membrane and cofactor routing, residue-level topology, glycosylation/disulfide risk, and a ranked binding-panel hypothesis.

The centrepiece of this paper is §3, where we walk every Astra AI model through one protease — Neprilysin / MME (P08473), a membrane-anchored zinc metalloendopeptidase targeted by sacubitrilat in sacubitril/valsartan that degrades amyloid-$\beta$ [8, 9]. In that worked example, Astra recovers the catalytic zinc context from sequence alone, calls the membrane-anchor state, identifies the glyco-rich construct-design problem, and returns a ranked ligand/cofactor panel for expert review.

The strongest use cases are where protease programs spend money: catalytic zinc and inhibitor-context review, membrane-anchor routing, glycosylation and disulfide construct signals, and family-level triage across the degradome. The buyer value is the ranked ligand panel and its residue-level pocket localization, which supports structural handoff. Exact PDB ligand-ID recovery is reported separately and deliberately as the strictest possible test — a single best-guess three-letter crystallographic component identifier matched against the full PDB ligand vocabulary (Recall@1 0.053; per-target set-F1 medians 0.29–

0.50 by catalytic class) — so it reads low by construction and is a governance check on panel identity, not the product promise. No experimental protease $\Delta T_m$ validation is claimed.

**What This Document Is.** A performance report on Orbion's Astra AI Suite on the human protease class. Every number traces to a checked-in source-data artifact. The classification, topology, and PTM metrics characterise how the deployed models behave on this class in production — including proteins drawn from the models' training corpora; held-out generalisation is reported in the per-model preprints [1–3]. Ligand binding is scored against PDB co-crystal contacts; exact three-letter ligand-ID recall is reported as the strictest governance check. No experimental protease $\Delta T_m$ validation is claimed (§6).

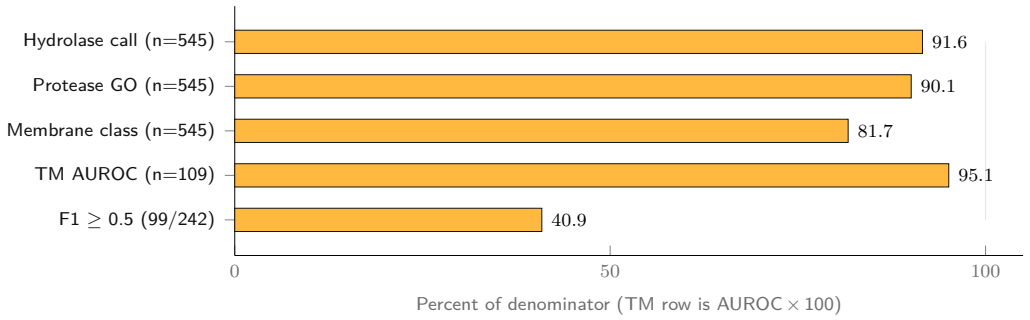



***Figure 1:*** *Protease suite signals that travel into a decision review. The binding row is restricted to the PDB-contact-scored subset; the other rows use the denominators shown in the labels.*

## 1 Why Proteases

Proteases are among the most clinically validated target classes in medicine: the ACE inhibitors and the neprilysin inhibitor sacubitril in cardiovascular disease, the proteasome inhibitors in oncology, the HIV, HCV, and SARS-CoV-2 protease inhibitors in antiviral therapy, and the DPP-4 inhibitors in metabolic disease all act on this class. Yet the hard targets still fail early for practical reasons: unstable constructs, ambiguous membrane context, shallow or moving substrate grooves, missing co-crystal guidance, and paralogue selectivity risk. The opportunity is not to claim that proteases are unexplored — it is to make the difficult fraction of the human degradome easier to route from sequence to a tractable first experiment.

The three risks that decide protease spend are concrete. **Pocket risk.** Catalytic clefts, metal sites, exosites, and substrate grooves must be prioritized before co-crystal and screening spend — a metalloprotease zinc cleft and a coagulation-protease exosite demand different chemistry. **Construct risk.** Membrane anchors, zymogen or prodomain state, glycosylation, and disulfides decide whether the assay measures the intended state; a type II single-pass membrane protease expressed as a naive soluble construct can lose the very context the program cares about. **Selectivity risk.** Paralogue families require early off-target thinking, not a late counter-screen after the chemistry funnel widens — the M1 aminopeptidases and the matrix-metalloprotease degradome are classic settings where one lead can engage a dozen relatives.

A useful AI output therefore has to say more than "there is a protease domain." It has to connect family, topology, cofactor, PTM, activation-state risk, and binding evidence into a decision package: which target is ready for experimental follow-up, what construct route is plausible, and which ligand or chemotype panel deserves the next co-crystal or

assay.

Orbion's Astra AI Suite uses one sequence input to return complementary outputs: role and family, membrane class, residue topology, PTM sites, and a ranked ligand-resolved binding panel (Figure 2). For a protease buyer, that means fewer blind construct variants, fewer undirected co-crystal attempts, and a defensible first active-site panel before expensive wet-lab routing begins.

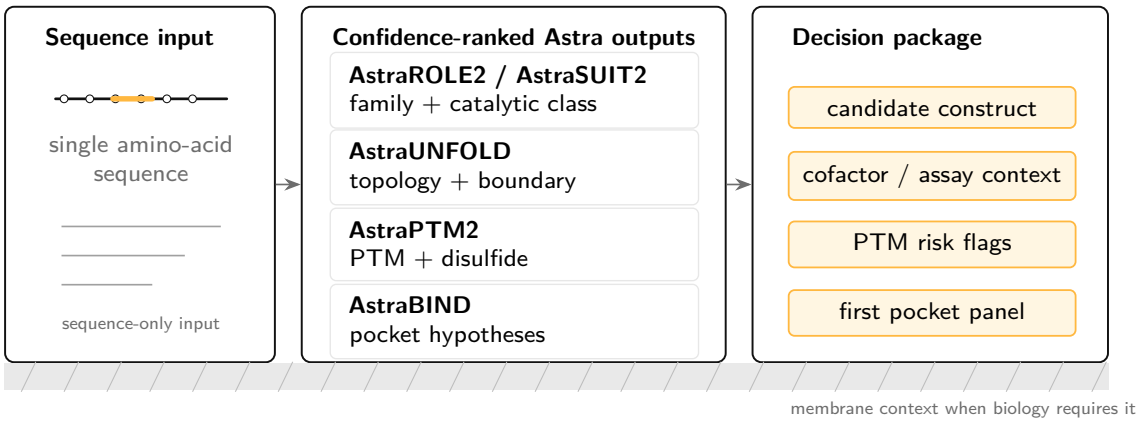



***Figure 2:*** *Astra maps a protease sequence to confidence-ranked decisions a program team can act on: candidate construct, assay context, PTM/disulfide risk, and first pocket panel.*

## 2 The Astra AI Suite, Capability by Capability

This volume evaluates five Astra prediction areas for protease work: AstraROLE2, AstraSUIT2, AstraUNFOLD, AstraPTM2, and AstraBIND. Each answers a different buyer question. Together they turn a sequence into a first-pass decision package for target triage, construct design, and binding-panel prioritization.

Thermostability is intentionally outside the claim set. The protease evidence available for this volume does not support an experimental $\Delta T_m$ validation statement, so topology and disorder outputs are used for construct and region triage only.

### 2.1 Protein Function and Protease Family

**What We Predict.** AstraROLE2 returns broad protein category, EC top class, GO molecular-function labels [6, 7], and pathway memberships from sequence alone [2], answering whether a sequence behaves like protease chemistry within a curated protease panel. Inside the 545-protein protease cohort, the operative readout is the hydrolase / protease signal rather than a single coarse category. MEROPS-style clan or subfamily assignment is not claimed in this volume.

**Headline Numbers.**

- Hydrolases (EC 3) are predicted on 499 of 545 eligible proteins (91.6 %).
- The protease / peptidase GO label is predicted on 491 proteins (90.1 %).
- The coarse "Enzymes" category is predicted on 526 proteins; dual identities such as immune, storage, regulatory, and membrane proteins explain much of the remainder.

**Strong and Weak.** Strong as a mechanism-routing signal that a sequence belongs in a protease workflow. Weak as a specificity claim: these numbers are enrichment within a

curated protease cohort, not a test against non-proteases, and the coarse protein-category head is the least reliable of the three outputs.

**Use as.** An intake signal for routing a sequence into protease-specific structure modeling, construct design, and binding-panel triage.

### 2.2 Topology, Cofactor, and Membrane Class

**What We Predict.** AstraSUIT2 returns soluble / peripheral / single-pass / multi-pass membrane class, TM-helix bin, localization, cofactors, and related structural categories, telling a program whether to build a soluble construct, a membrane-retaining construct, or both.

**Headline Numbers.**

- Top membrane-class prediction matches UniProt TM-feature-derived class on 445 of 545 eligible proteins (81.7 %).
- Metal-ion cofactor is predicted on 156 proteins, with zinc specifically predicted on 124.

**Strong and Weak.** Strong for membrane-class routing and cofactor presence, including zinc — membrane proteases fail differently from soluble proteases. Weak as a determination of whether a predicted cofactor is catalytic or structural; that call requires the residue-level and binding evidence below.

**Use as.** Route with AstraSUIT2, then use AstraUNFOLD for residue-level construct boundaries when an anchor or ectodomain decision affects expression spend.

### 2.3 Residue-Level Topology and Disorder

**What We Predict.** AstraUNFOLD returns per-residue membrane, inside/outside, disorder, and amyloidogenicity probabilities, locating where a construct boundary should move when topology matters. Against UniProt transmembrane annotations, residue-level membrane probability reaches mean AUROC 0.951 and mean AUPRC 0.804 [12] on 109 annotated proteases; median AUROC is 0.995 (Figure 3).

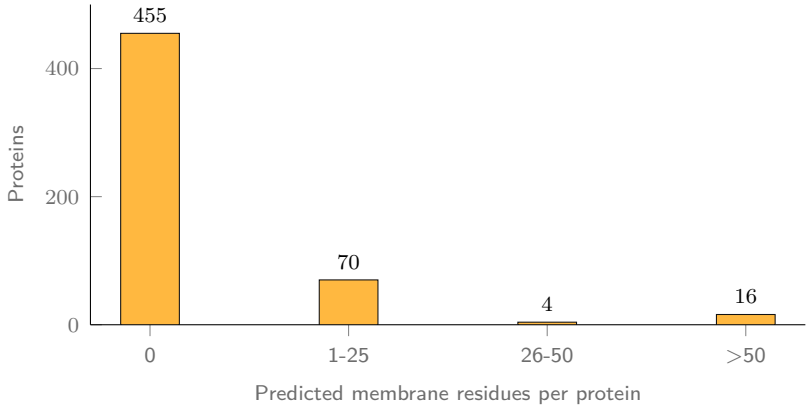



***Figure 3:*** *Predicted membrane-residue counts across eligible proteases. The non-zero tail is the construct-design queue for membrane-anchored targets.*

**Strong and Weak.** Strong at recovering annotated transmembrane spans for anchor and ectodomain boundary decisions. Weak as a substitute for experimental topology: the AUROC is computed on the 109 proteases carrying UniProt transmembrane annotations, whereas the histogram above counts predicted membrane residues across all 545 eligible proteins, so its non-zero tail (90 proteins) need not equal the annotated set.

**Use as.** Preserve anchors when biology requires a membrane context, or design ectodomain constructs with the anchor removed when the assay allows it.

### 2.4 Post-Translational Modification Site Prediction

**What We Predict.** AstraPTM2 emits high-precision and high-recall PTM site sets across 39 modification classes [1], flagging which sites affect expression host, construct boundaries, and mutation-avoidance choices. The two operating points serve different jobs: the high-precision call is for confident wet-lab handoff (few false positives), and the high-recall call is for exhaustive hypothesis generation; below we cite whichever point is operative for each modification and label it. For protease programs, structural modifications are the most immediate commercial signal.

**Headline Numbers.**

- N-linked glycosylation reaches high-recall F1 0.924 against UniProt feature annotations.
- Disulfide bonds reach high-precision F1 0.829.
- Phosphorylation is weaker as a thresholded exact-site task (high-precision F1 0.149) and should be treated as hypothesis-generating.

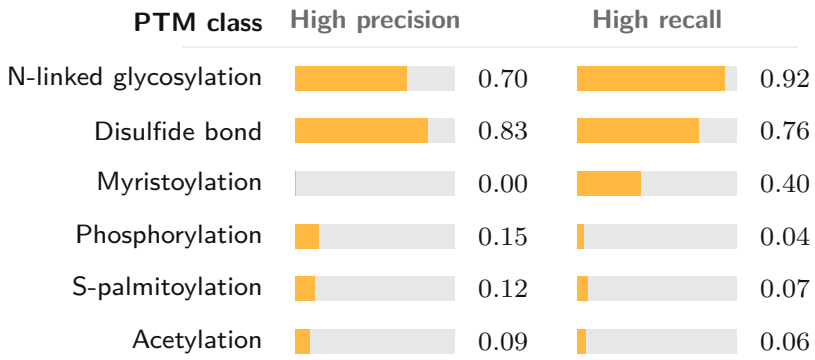



***Figure 4:*** *Exact-position PTM F1 against UniProt feature annotations. Structural modifications dominate the construct-design signal.*

**Strong and Weak.** Strong for the structural modifications that drive construct design — N-linked glycosylation and disulfide bonds (Figure 4). Weak for phosphorylation as a thresholded exact-site task, which remains a follow-up hypothesis rather than a definitive regulatory map.

**Use as.** For secreted and membrane proteases, glycosylation and disulfide calls inform expression host choice, domain boundaries, and mutation avoidance.

### 2.5 Ligand-Resolved Binding Panel

**What We Predict.** AstraBIND returns a ranked ligand panel and candidate contact residues from sequence alone [3], surfacing which ligand or cofactor hypotheses to review first for a PDB-contact-scored protease. This volume scores exact PDB ligand-ID recovery against 4.0 Å heavy-atom contacts in the subset of proteins with suitable co-crystal evidence.

**Headline Numbers.**

- Strict ligand-identity scoring covers 242 PDB-contact-scored proteins and 2,396 (protein, ligand) ground-truth pairs, not all 545 eligible proteins.
- Set-F1 (micro) is 0.367 (precision 0.419, recall 0.326).
- 99/242 scored proteins reach set-F1 $\geq 0.5$, defining a higher-confidence review subset for ligand-panel follow-up.
- Recall@1 is 0.053, Recall@3 0.137, Recall@5 0.200, and mean reciprocal rank 0.119; these are exact ligand-ID retrieval numbers, separate from residue-level pocket localization.
- Residue-level scoring covers 833 matched ligand-residue panels and 242 proteins with contact residue ground truth. Median ligand-residue F1 is 0.389 and median per-residue AUROC is 0.660.

**How to use the threshold.** The set-F1 $\geq 0.5$ bucket is an operating triage threshold: it marks proteins where the recovered ligand-ID set is rich enough to justify expert panel review. It is not a binary success claim, and proteins below the threshold can still be useful when the recovered item is a cofactor or a mechanistically informative inhibitor-like ligand. The headline set-F1 0.367 is the micro-pooled corpus figure; the 99/242 count and the per-target examples are per-protein set-F1, so they are not directly comparable to the pooled value.

**Representative binding-panel behavior.**

| Protease | Recovered set | Top-five panel IDs | Decision meaning |
|---|---|---|---|
| Neprilysin / MME | 11/14; set-F1 0.85 | ZN, 6LD, BIR, D0W, FT8 | Metal cofactor plus structural ligand IDs: upper-tail positive control for zinc-cleft review. |
| Caspase-1 / CASP1 | 11/16; set-F1 0.79 | MG, 158, 3NS, 4QB, 5PH | Structural/context review bucket: high-recovery panel for co-crystal triage. |
| MMP-2 | 4/4; set-F1 0.50 | 2HP, B3P, CA, I52, L2U | Structural/context bucket: all four observed IDs are recovered, but set-F1 0.50 reflects low precision (0.33) — most predicted IDs are not observed, and additive-like components can rank in the raw panel. |
| Thrombin / F2 | 5/136; set-F1 0.068 | 2CE, 4CP, 896, 897, ABN | Hard-inventory bucket: useful hits can appear, but broad co-crystal diversity depresses strict recall. |

**Strong and Weak.** Strong as a ranked hypothesis generator: exact ligand-ID recall governs panel identity, and residue-level F1 and AUROC show whether predicted contacts localize around the observed pocket (median recovery by catalytic class in Figure 5).

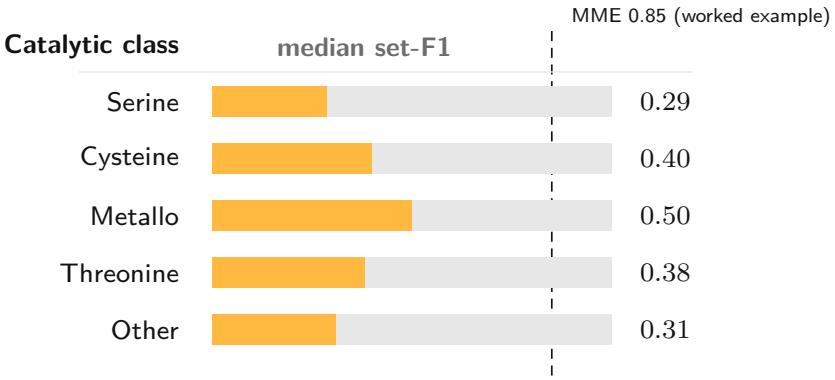



***Figure 5:*** *Catalytic-class-label view of median binding-panel recovery in the PDB-contact-scored subset. Rows are descriptive labels, not additive denominator buckets; the global binding denominator remains 242 scored proteins. Per-class scored n: Serine 66, Cysteine 66, Metallo 72, Threonine 7, Other 32 — the Threonine median rests on a small sample.*

Weak on strict exact-ID recall (Recall@1 0.053), which is mechanically low, and on hard-inventory targets such as thrombin where broad co-crystal diversity depresses recall.

**Use as.** AstraBIND is best used as a ranked hypothesis generator for ligand-panel review, cofactor context, and follow-up structural or biochemical design.

**Safe binding-panel workflow.**

1. Freeze the ranked panel at the start of a decision cycle so chemists, structural biologists, and program leads review the same candidate set.
2. Separate cofactors, inhibitor-like ligands, substrate fragments, and context ligands before interpreting any single PDB identifier as a drug-like chemotype.
3. Use exact-ID recovery as governance, then inspect candidate contacts and known catalytic or metal-binding residues before selecting co-crystal, competition, or counter-screen experiments.

## 3 Worked Example: Neprilysin Across the Full Suite

Aggregate metrics are useful for governance; buying decisions happen one target at a time. Neprilysin (MME, UniProt P08473) was selected as a transparent case study because it is commercially relevant, membrane anchored, zinc dependent, glycosylated, disulfide rich, and co-crystal supported. That makes it a useful positive-control target, not a claim that every protease will behave the same way.

MME is a 750-aa type II single-pass membrane M13 gluzincin metalloendopeptidase. It is inhibited by sacubitrilat in sacubitril/valsartan for heart failure [8] and degrades amyloid-$\beta$ [9]. In the PDB-contact-scored binding subset, its set-F1 of 0.85 ranks it among the top 18 of 242 scored proteases (Figure 6). Exact top-1 ligand-ID recall across the cohort is low (Recall@1 0.053); the value on this target is zinc-cleft localization and panel triage, not first-pick ligand identity. Realistic per-target expectation sits at the catalytic-class medians (set-F1 0.29–0.50), so Neprilysin is an upper-tail positive control, not the typical case.

**AstraROLE2.** The function head calls Hydrolases (EC 3) with score 0.988 and protease / peptidase GO activity with score 0.950. That routes the sequence to protease chemistry rather than a generic membrane-protein workflow.

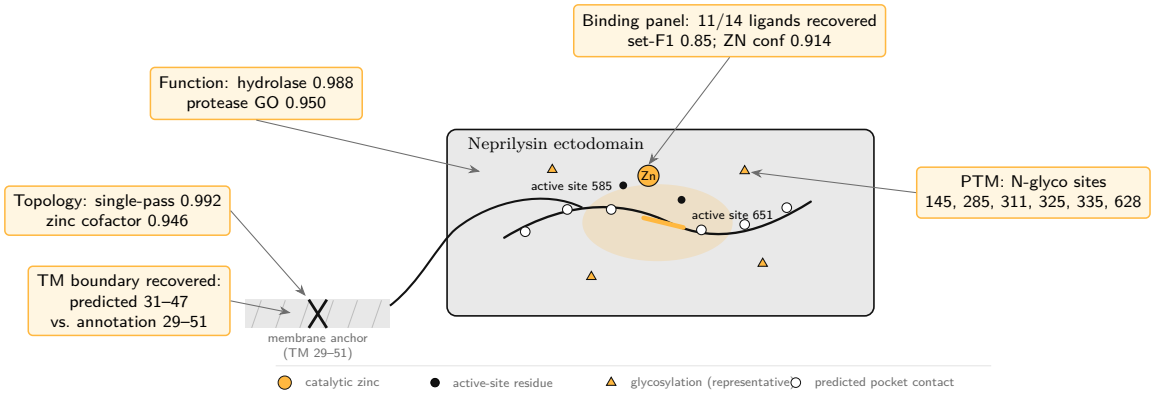



***Figure 6:*** *Integrated sequence-only Astra output for Neprilysin (gene MME), a membrane-anchored zinc metalloprotease of the M13 / gluzincin family. The central cartoon is the protein; each amber callout is one Astra model output mapped onto the feature it describes — function (AstraROLE2), membrane class (AstraSUIT2), transmembrane span (AstraUNFOLD), modification sites (AstraPTM2), and the ranked binding panel at the catalytic cleft (AstraBIND). See the key for the cartoon symbols.*

**AstraSUIT2.** The topology and biochemical head calls **Single-pass** membrane protein at score 0.992 and **1 TM Helix** at score 0.992, matching the type II signal-anchor annotation at residues 29–51. It also calls zinc cofactor with score 0.946. The program implication is direct: design the ectodomain and membrane-retaining construct routes intentionally.

**AstraUNFOLD.** The residue-level head predicts the membrane signal across residues 31–47, inside the annotated TM span. That gives a construct-boundary hypothesis from sequence alone.

**AstraPTM2.** The high-recall PTM output predicts N-linked glycosylation at 145, 285, 311, 325, 335, 628, recovering the annotated glycosylation positions 145, 285, 325, 628 and adding two hypothesis-grade ectodomain sites. The high-precision disulfide output flags cysteine positions 57, 62, 80, 88, 143, 234, 621, 695, 735, 747. That informs expression host, glyco-engineering, and mutation-avoidance decisions before construct ordering; it is not a disulfide-pair topology assignment.

**AstraBIND.** Neprilysin has 14 PDB co-crystal ligand IDs in the strict ligand-identity benchmark. AstraBIND predicts 12 ligand IDs and recovers 11 of the 14, for set-F1 0.85. The frozen top-five panel is ZN, 6LD, BIR, D0W, and FT8: ZN is the metal-cofactor bucket (ranked at panel confidence 0.914), and the other top IDs are structural-panel ligand identities at the cleft. These identities include inhibitor-like PDB components and context ligands; they should be reviewed as a ranked panel, not collapsed into a single drug-like chemotype claim.

The zinc-site cross-check is deliberately conservative. UniProt annotates active-site residues 585 and 651 and metal-binding residues 584, 588, 647. AstraBIND's ZN prediction includes the annotated active-site residues 585/651 and the metal-binding residues 584, 588, and 647 that form the zinc-coordination shell of the M13 gluzincin active site (the active-site residues are catalytic, not direct zinc ligands). UniProt's residue 103 is a substrate (peptide) binding site, not a metal ligand, and is outside the predicted ZN set; model-only residues are not re-described as catalytic residues.

The residue handoff is also explicit. The PDB-contact ground truth for MME ZN contains residues 584, 588, 647, 712; AstraBIND's ZN panel marks 584, 585, 586, 588,

591, 647, 651, 712, recovering all four contact residues (residue recall 1.00, residue F1 0.667 at residue precision 0.50). Across all MME binding-panel entries, AstraBIND localizes predicted contacts to a focused set spanning the ectodomain — including all four PDB zinc-contact residues (584, 588, 647, 712) at the catalytic cleft — with target-level per-residue AUROC 0.632. This is the pocket-localization evidence a structural biologist can inspect before choosing the first co-crystal or competition experiment.

**Neprilysin structural handoff.**

| Decision | Verified MME evidence | Program action |
|---|---|---|
| Construct state | Type II single-pass membrane annotation 29–51; AstraUNFOLD membrane signal 31–47. | Prepare soluble ectodomain and membrane-retaining routes deliberately; do not treat the sequence as a generic soluble protease. |
| Cofactor anchor | Zinc cofactor score 0.946; ZN prediction recovers PDB-contact residues 584, 588, 647, 712 and overlaps active-site residues 585/651. | Anchor structural inspection around the zinc cleft and preserve metal-site chemistry in assay design. |
| PTM and expression | Predicted N-linked glycosylation sites 145, 285, 311, 325, 335, 628; annotated positions 145, 285, 325, 628 are recovered; high-precision disulfide output flags cysteine positions 57, 62, 80, 88, 143, 234, 621, 695, 735, 747. | Choose expression host and mutation-avoidance plan before construct ordering; preserve flagged cysteines during trimming or mutagenesis, and do not infer pair topology without structural or biochemical evidence. |
| Binding panel | 11/14 strict co-crystal ligand IDs recovered; set-F1 0.85; top-five IDs ZN, 6LD, BIR, D0W, FT8; target-level pocket AUROC 0.632. | Split cofactors, inhibitor-like components, and context ligands before selecting co-crystal, competition, or analogue-panel experiments. |
| Selectivity | MME and ECE1 return frozen, separable ligand and residue-contact panels (MME 11/14, pocket AUROC 0.632; ECE1 2/2 at precision 0.25, pocket AUROC 0.805); MMEL1 returns a ranked ligand panel from sequence with no co-crystal ground truth to score. | Run the same frozen panel format across the scored nearest comparators (MME, ECE1) and review MMEL1's predicted panel before widening the screening funnel. |

**Positioning against common alternatives.** Structure predictors and template-transfer tools are useful in protease work, but they answer a different question: fold or complex modeling, often one target or ligand at a time [10, 11]. AstraBIND's role in this workflow is the ranked ligand-panel hypothesis from sequence alone. This paragraph is a capability contrast, not a head-to-head accuracy benchmark.

### What a Program Team Does With This Output

- Treat the recovered zinc cleft as the orthosteric anchor for a metalloprotease-inhibitor campaign.
- Review the recovered ligand identities for co-crystal, competition, or analogue-panel design rather than starting from an unranked apo pocket.
- Design constructs around the type II signal-anchor: ectodomain-only for soluble expression where acceptable, membrane-retaining when the assay needs native context.
- Preserve or engineer around the glycosylation and disulfide map before expression-host selection.
- Use the same frozen panel format across paralogues to triage selectivity and off-target degradome risk before the screening funnel widens.

This is not a claim that the model replaces structural biology. It is a claim that a protease program can decide what structural biology to buy first.

## 4 Limits and Operating Envelope

Five boundaries define the production envelope for this protease volume.

**Exact ligand-ID recovery is stringent.** Recall@1 is 0.053 on 2,396 exact (protein, ligand) ground-truth pairs in the PDB-contact-scored subset. This task asks whether the ranked panel contains the exact PDB ligand ID observed in a co-crystal. It should not be confused with chemotype usefulness, cofactor recovery, or residue-level pocket localization.

**Binding panels should be frozen for decision review.** The binding head can produce run-to-run differences in ranked panel details. For buyer work, a panel should be generated, reviewed, and frozen at the start of a decision cycle so chemistry, structural biology, and program leadership are discussing the same candidate set.

**No experimental protease $\Delta T_m$ claim.** This volume covers five models and makes no sign-accuracy, correlation, or independent-generalization claim for protease thermostability. The bounded value is construct and region triage from topology/PTM outputs, not a validated protease $\Delta T_m$ predictor.

**Residue-level pocket evidence is model evidence.** The protease results include residue-level pocket scoring: median ligand-residue F1 0.389 across 833 matched panels, and median per-residue AUROC 0.660 across 242 proteins. These numbers support structural handoff; they are not prospective wet-lab validation.

**Exosites and diffuse substrate grooves are harder.** Compact catalytic metal or inhibitor pockets are where the current binding head is most useful. Exosite binders, glycan-mediated contacts, shallow substrate grooves, and extracellular-domain-only interactions should be routed to structural inspection after the sequence-only panel is generated.

## 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 protease program questions to the prediction-area combinations that address

them and the expected output for a discovery or protein-engineering team. One sequence-input call gives a program team enough complementary evidence to prepare a decision review instead of commissioning five disconnected analyses.

***Table 1:*** *Decision framework mapping common protease 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. Novel or Orphan Protease** | Is this target ready for catalytic-pocket follow-up? Run **AstraROLE2** + **AstraSUIT2** + **AstraUNFOLD** + **AstraBIND** to return an active-site hypothesis, cofactor context, a ranked ligand panel, and a construct route for validation. |
| **2. Selectivity Risk** | Which paralogues need early counter-screening? Run **AstraBIND** + **AstraROLE2** across the target family to produce a shared/divergent ligand-panel shortlist for degradome off-target review. |
| **3. Membrane-Protease Construct** | Should the program build soluble, membrane-retaining, or parallel constructs? Run **AstraSUIT2** + **AstraUNFOLD** + **AstraPTM2** to return expression-ready construct options with anchor, ectodomain, glycosylation, and disulfide risks. |
| **4. Ligand-Panel Prioritization** | Which ligand classes deserve the next co-crystal, competition assay, or analogue panel? Run **AstraBIND** with **AstraSUIT2** / **AstraROLE2** context to return ranked binding-panel hypotheses with explicit caution for exact ligand-ID recovery. |
| **5. PTM and Regulatory-Site Map** | Which modification sites should guide expression-host choice, MS follow-up, or mutation avoidance? Run **AstraPTM2** filtered by **AstraUNFOLD** topology/disorder and **AstraSUIT2** localization to return a modification map prioritized for construct design and follow-up analytics. |

The recurring pattern across all five workflows is the same: a primary prediction area produces a candidate set, and one or two complementary areas filter or contextualize the candidates against other protein properties. 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.

**What the protease evidence pack looks like.**

| Decision layer | Buyer-ready output |
|---|---|
| Target triage | Lead target and comparator proteins, mechanism class, confidence tier, and the specific reason each sequence is ready or not ready for wet-lab follow-up. |
| Protease state and construct | Membrane-retaining versus ectodomain route, activation/prodomain risk flag, glycosylation and disulfide liabilities, and mutation-avoidance notes for construct ordering. |
| Binding-panel review | Cofactors, inhibitor-like ligands, substrate/context ligands, and exact-ID misses separated before co-crystal, competition, or analogue-panel planning. |
| Selectivity screen | Nearest paralogue and degradome counter-screen shortlist, with shared versus divergent binding-panel hypotheses highlighted for medicinal chemistry review. |
| Assay handoff | The first structural or biochemical experiment to buy, plus the caveat that prospective wet-lab validation and protease $\Delta T_m$ validation are outside this volume. |

**Comparator-panel selectivity examples.**

| Counter-screen set | Observed Astra evidence | Buyer use |
|---|---|---|
| Neprilysin family | MME recovers 11/14 strict ligand IDs (set-F1 0.85) with pocket AUROC 0.632. ECE1 recovers 2/2 IDs (set-F1 0.40 on a small 2-ligand panel) with strong pocket localization (AUROC 0.805; precision 0.25 — the 2-ligand ground truth makes recall easy but most of the 8 predicted IDs are unobserved, so read the panel by type). MMEL1 returns a ranked ligand panel from sequence (predicted, unscored, no co-crystal ground truth). | Put the scored comparators MME and ECE1 into the first degradome counter-screen, with MMEL1's predicted panel as an unscored add-on, before expanding chemistry. |
| M1 aminopeptidases | ERAP2 recovers 10/11 IDs (set-F1 0.87), LNPEP 8/9 (0.76), and ERAP1 3/8 (0.30), with divergent top panels. | Prioritize family counter-screens by panel overlap, recovery tier, and pocket-localization confidence. |
| MMP slice | Ligand-ID recovery is upper-to-mid for MMP1/7/8, mid for MMP2, and lower for MMP3/9; per-residue pocket AUROC is comparatively stable (0.718–0.851) across all six, so localization holds even where exact-ID recovery drops. | Select the smallest informative MMP counter-screen set instead of treating the degradome as undifferentiated. |

## 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, topology and cofactor classification, post-translational modification site prediction, residue-level topology/disorder, and ligand binding-site identification.

For protease programs, the decision package is concrete: target shortlist, construct-

design route, ranked ligand-panel hypotheses, selectivity counter-screen list, and an evidence pack that a discovery team can take into wet-lab planning.

> **What the reader should remember.** Astra is not positioned as a replacement for structural biology or protease biochemistry. It is a front-end decision layer: which construct route to prepare, which cofactor or ligand hypotheses to inspect, which paralogues to counter-screen, and which wet-lab experiment to buy first.

**Start with one target.** Send a single protease sequence — an orphan target, a difficult membrane protease, or a paralogue you need to counter-screen — and Orbion returns the full sequence-only decision package described above for expert review, with the operating envelope stated up front so your team knows exactly what to trust and what to test.

**Platform.** app.orbion.life **Web.** orbion.life **Contact.** `contact@orbion.life`

## 6 Methods

**Cohort.** The evaluation cohort is 546 reviewed human proteases from UniProt Swiss-Prot [4]; the all-organism reviewed protease universe contains 12,920 rows. One sequence failed a length eligibility check, leaving 545 eligible proteins.

**Reference Annotations.** Sequence-level reference annotations come from UniProt feature tables: transmembrane segments, active sites, binding sites, modified residues, glycosylation sites, disulfide bonds, domains, and regions. PDB ligand-contact reference uses heavy-atom contacts at 4.0 Å between ligand and protein residue, aggregated per UniProt accession across deposited co-crystal structures [5]. Buffer, cryoprotectant, and common-additive ligands are excluded before scoring.

**Reported Metrics.** Role and Suit are reported as cohort-level routing signals. Unfold is reported against UniProt transmembrane residue annotations. PTM is reported as exact-position F1 against UniProt feature annotations. Binding is reported on the PDB-contact-scored subset: 242 proteins and 2,396 (protein, ligand) ground-truth pairs. Residue- and site-level agreement is reported as AUROC (area under the receiver-operating-characteristic curve), AUPRC (area under the precision–recall curve), and F1 (the harmonic mean of precision and recall) [12].

**Binding Interpretation.** Recall@k asks whether a PDB-observed ligand ID appears in the top-k ranked AstraBIND panel. MRR assigns zero when the ligand is absent. Set-F1 compares unordered predicted and observed ligand-ID sets. These metrics are strict ligand-identity metrics; they do not by themselves establish residue-level pocket recovery. Per-residue AUROC is computed per target with the observed 4.0 Å contact residues as positives and the remaining sequence positions as negatives. The additive exclusion above applies to the scored reference set; the ranked panels shown for individual targets are the model's raw output and can still surface additive-like or cofactor identities, which is why panels are read by type rather than as drug-like chemotypes.

**Generalization.** The classification, topology, and PTM cohort is drawn from the reviewed proteome, which overlaps the corpora these models were trained on, so those figures characterise within-distribution behaviour on real targets rather than held-out generalisation. The binding numbers are recovery figures against deposited co-crystal contacts; held-out generalisation is not separately claimed in this volume.

**Score Interpretation.** Model probability and confidence values in the case study are reported as model output scores. Calibration of those scores is not evaluated in this

volume.

**Clinical and Scientific References.** External literature is used for biological context, benchmark interpretation, and method comparison only. The Neprilysin case study cites clinical and biology references for target relevance; it is not presented as a clinical outcome prediction.

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

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

[8] McMurray, J. J. V. *et al.* Angiotensin-neprilysin inhibition versus enalapril in heart failure. *New England Journal of Medicine*, **371**:993–1004 (2014). doi:10.1056/NEJMoa1409077.

[9] Iwata, N. *et al.* Metabolic regulation of brain Abeta by neprilysin. *Science*, **292**(5521):1550–1552 (2001). doi:10.1126/science.1059946.

[10] Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, **630**:493–500 (2024). doi:10.1038/s41586-024-07487-w.

[11] Hekkelman, M. L. *et al.* AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nature Methods*, **20**:205–213 (2023). doi:10.1038/s41592-022-01685-y.

[12] 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.