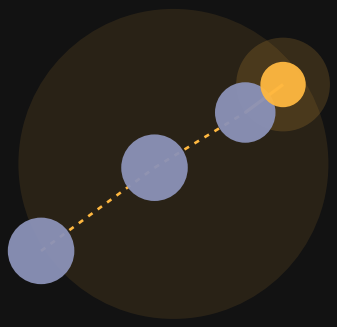
orb::ion



Model Performance Series

# Astra AI on Enzymes

How Orbion's six AI models perform across 6,304 catalytic proteins.


AUTHORS

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

[1]Orbion GmbH, Berlin, Germany

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



June 9, 2026 · orbion.life

## Executive Summary

Enzymes are the catalysts of biology and the most heavily exploited drug-target class in the history of medicine — the kinase inhibitors of oncology, the protease inhibitors of antiviral therapy, and the metabolic-enzyme blockers behind the statins and the gliptins all target this class. They are defined not by a shared fold but by the chemistry they perform, organised into the seven top-level classes of the Enzyme Commission. Three properties make enzymes a distinctive computational target: function is the defining question (which reaction, not which shape); catalysis localises to a handful of active-site residues; and the class is predominantly soluble rather than membrane-embedded. Each shapes how a useful model must behave.

This document reports how Orbion's Astra AI Suite performs across six prediction areas on the enzyme class. We ran every Astra AI model on **6,304 reviewed enzyme-associated proteins** 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], PDB co-crystal contacts at 4 Å resolution for ligand binding [5], and **644 mutations across deeply-scanned enzymes** with curated experimental thermal-shift data for stability prediction [9, 10].

**Headline Numbers Across Six Prediction Areas on Enzymes**

- **Function and EC Classification.** 95.3 % of the cohort is recognised as enzymatic via the EC head, distributed across all seven EC classes in proportions that track the enzyme universe (transferases and hydrolases dominate). Molecular-function GO accuracy: top-5 94.2 %, top-1 84.0 %.
- **Topology and Membrane Class.** The class is predominantly soluble — ~78 % carry no transmembrane segment — and the model routes them as such rather than imposing a membrane template.
- **Residue-Level Topology.** On the 1,345 membrane enzymes, per-residue prediction agrees with UniProt-annotated transmembrane segments at AUROC 0.95 (median 0.99), uniformly across all seven EC classes (0.89–0.98).
- **Post-Translational Modification (PTM) Site Prediction.** On the strongest classes: F1 0.89 for myristoylation, 0.87 for N-linked glycosylation, 0.86 for disulfide bonds. Phosphorylation — the dominant enzyme modification (11,474 sites across 2,616 enzymes) — is ranked at AUROC 0.95. All 39 classes covered; two operating points reported per class.
- **Ligand Binding Pocket Prediction.** 62 % ligand-identity recall and 82 % pocket-success across 2,318 enzymes with co-crystal data; strongest on the chemically well-defined cofactor and nucleotide pockets (iron / 2-oxoglutarate dioxygenases, kinase ATP sites, FAD/NAD flavoenzymes; residue F1 up to 1.00).
- **Thermostability Prediction ($\Delta T_m$).** The flagship result of the suite. On a deep scan of bacteriophage T4 lysozyme — the canonical protein-stability reference — Spearman $\rho$ 0.93 (n=315); across the full enzyme set, $\rho$ 0.88 with 90 % directional accuracy on strong-effect mutations.

Enzymes are where the suite is strongest, and the reason is structural. Function classification leads with the EC head, which is the right tool for a class defined by chemistry;

topology prediction succeeds by recognising that most enzymes are soluble; binding-pocket prediction is carried by the chemically distinct cofactor sites that define catalytic centres; and thermostability prediction — the single most valuable capability for enzyme engineering — reaches the strongest correlation we have measured anywhere in the suite. Where the model is weaker (rare PTM classes, the broad protein-category head, extreme-effect $\Delta T_m$ outliers), we report it directly in §4 rather than averaging it away.

The centrepiece of this paper is §3, where we walk every Astra AI model through one enzyme — the tyrosine kinase FYN (UniProt P06241), a Src-family signalling enzyme and oncology drug target — and show what a program team would do with the integrated output.

**What This Document Is.** A performance report on Orbion's Astra AI Suite on the enzyme 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 their training corpora; held-out generalisation is reported in the per-model preprints. Thermostability ($\Delta T_m$) is scored against curated experimental thermal-shift data; because the benchmark proteins are canonical stability systems, this measures performance on well-studied targets rather than certified held-out generalisation (§6).

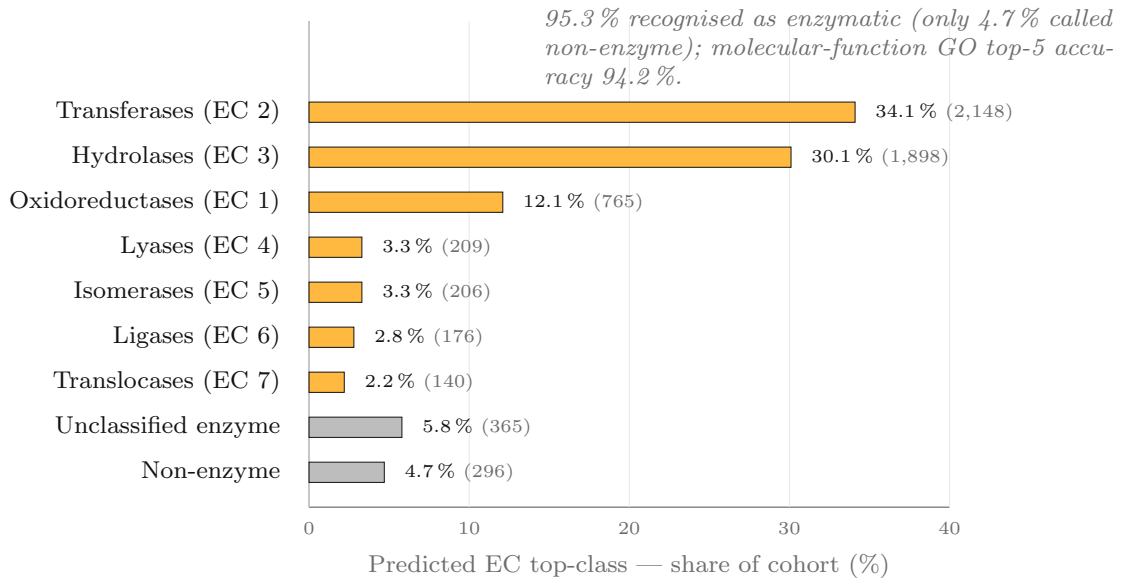



***Figure 1:*** *Function recognition across the enzyme cohort. The EC head recognises 95.3 % as enzymatic and distributes them across all seven EC top-classes (amber) in proportions that track the composition of the enzyme universe — transferases and hydrolases dominate, as they do in nature. Detail in §2.1.*

## 1 Why Enzymes

Enzymes are the catalysts of biology. They run metabolism, replicate and repair the genome, transduce signals, and degrade and remodel every other class of molecule in the cell. They are also the most heavily exploited drug-target class in history: kinase inhibitors anchor modern oncology, protease inhibitors transformed HIV and hepatitis-C therapy, and the statins, ACE inhibitors, and DPP-4 inhibitors are all enzyme blockers.

Where a receptor or transporter is a single point of intervention, an enzyme is a turnover machine — inhibit it and the catalytic output stops — which is why enzymes have been, and remain, the workhorses of small-molecule drug discovery. They are equally central to industrial biotechnology, where engineered hydrolases, transferases and oxidoreductases power detergents, diagnostics, food processing, and pharmaceutical manufacturing.

**Function is the defining property.** Unlike the receptor and transporter classes, enzymes are not grouped by a shared architecture — they are grouped by the reaction they catalyse. The Enzyme Commission (EC) hierarchy partitions all of enzymology into seven top-level classes — oxidoreductases (EC 1), transferases (EC 2), hydrolases (EC 3), lyases (EC 4), isomerases (EC 5), ligases (EC 6) and translocases (EC 7). A useful annotation model for this class must therefore lead with *function*: not "what fold is this" but "what chemistry does it do." This is a fundamentally different question from the one a topology model asks, and it is the question the Astra suite's function head answers directly (§2.1).

**Catalysis localises to a handful of residues.** An enzyme of three hundred residues may carry its entire catalytic power in three — the serine–histidine–aspartate charge-relay triad of the proteases is the textbook case, but metal centres, cofactor-binding motifs, and oxyanion holes follow the same principle: a small, precisely arranged constellation of residues does the chemistry while the rest of the chain is scaffold. This concentration of function makes the active site the single most important feature to locate, and the hardest, because it is defined by three-dimensional geometry rather than by sequence proximity. Pocket and binding-site prediction (§2.5) speaks directly to this.

**Most enzymes are soluble.** Where GPCRs are uniformly seven-transmembrane and transporters are uniformly polytopic, enzymes are predominantly soluble, globular proteins — in our cohort roughly four in five carry no transmembrane segment at all. The topology question therefore inverts: the task is not "where are the membrane-spanning helices" but "is this protein membrane-associated at all, and if not, the residue-level disorder and aggregation behaviour of its soluble fold." A model that assumes a membrane architecture would fail this class; we show in §2.2 that the suite routes soluble and membrane enzymes correctly rather than imposing one template on both.

**Stability is the engineering bottleneck.** For no other class is thermal stability so directly the currency of the field. Industrial biocatalysts are selected and engineered for the stability that lets them survive process conditions; therapeutic enzymes are optimised for shelf-life and serum half-life; and structural and mechanistic studies routinely depend on stabilising mutations to yield tractable protein. The mutational landscape is far too large to screen exhaustively by thermal-shift assay, and a computational prefilter that ranks candidate mutations by predicted $\Delta T_m$ is of immediate practical value. This is the flagship result of this report (§2.6): on enzyme mutations scored against curated experimental thermal-shift data, the model achieves the strongest stability correlation we have measured anywhere in the suite.

**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 — so that regularities learned across the broader proteome (catalytic motifs, cofactor signatures, fold-stability relationships) transfer to enzymes without per-family specialisation. The remainder of this document reports how the suite performs on the 6,304 reviewed enzyme-associated proteins in UniProt Swiss-Prot [4], one prediction area at a time, with the experimental reference

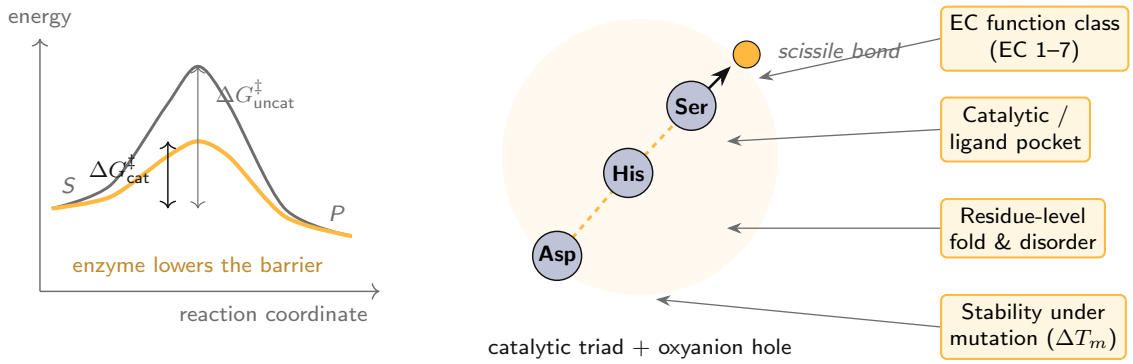



***Figure 2:*** *The catalytic logic that defines the enzyme class. An enzyme lowers the activation barrier of a reaction (left) through a small constellation of active-site residues — here the serine–histidine–aspartate charge-relay triad and oxyanion hole of a hydrolase (centre) — while the bulk of the chain is folded scaffold whose stability sets whether the catalyst survives at all. The prediction tasks the Astra suite addresses — the EC function class, the catalytic and ligand-binding pocket, the residue-level fold behaviour, and the stability of the fold under mutation — are annotated at right.*

data and metrics for each.

## 2 The Astra AI Suite, Capability by Capability

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 6,304-protein enzyme 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 [6, 7], and experimentally measured thermal shifts for stability. Methodological details, metric definitions, evaluation scope, and per-area 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. For binding-site prediction this property is reinforced by a seven-gate anti-hallucination audit (§2.5).

### 2.1 Protein Function and EC Classification

**What We Predict.** From a protein sequence alone, the protein's Enzyme Commission (EC) class, its broad functional category, its molecular-function Gene Ontology terms, and its pathway memberships. For enzymes the EC head is the operative output: it answers the defining question of the class — which of the seven reaction types does this protein catalyse. Evaluated on all 6,304 proteins in the enzyme cohort.

**Headline Numbers.**

- **Enzyme recognition:** **95.3 %** — the EC head assigns an enzymatic class to all but 4.7 % of the cohort, with a mean enzyme probability of 0.95.
- **All seven EC classes resolved** in biologically sensible proportions: transferases 34 %, hydrolases 30 %, oxidoreductases 12 %, then lyases, isomerases, ligases and

translocases (Figure 1). This tracks the known composition of the enzyme universe, where transferases and hydrolases dominate.

- **GO molecular-function (ancestor closure): top-5 94.2 %**, top-1 84.0 %.

**Strong and Weak.** The coarse protein-*category* head, by contrast, labels only 68.8 % of the cohort "Enzymes" — not a miss, but a reflection of dual identity: a receptor tyrosine kinase is also a signalling protein, a regulatory protease also a regulatory protein. For a class defined by chemistry, the EC head is the correct readout, and it is the one to cite (§4). The predicted EC-class distribution is itself a consistency check: with no per-protein EC labels supplied, the model reproduces the field's known class proportions.

**Use as:** production triage on enzymatic function — EC-class assignment and molecular-function annotation for uncharacterised or hypothetical enzymes.

### 2.2 Topology and Membrane Class

**What We Predict.** Membrane topology class (soluble / peripheral-membrane / single-pass / multi-pass), transmembrane-helix count, and auxiliary biochemical classifications (cofactor, subcellular localization, quaternary structure).

**Headline Numbers.** The enzyme cohort is predominantly soluble, and the model reflects that rather than defaulting to a membrane template: **62 % soluble**, 15 % peripheral-membrane, 13 % single-pass, 9 % multi-pass. Soluble and peripheral together — the ∼78 % with no transmembrane segment — match the fraction the residue-level head independently routes as non-membrane.

**Strong and Weak.** This is the topological inverse of the receptor and transporter classes. The value here is correct routing: the model does not hallucinate transmembrane helices into the soluble majority, which is what lets the residue-level head below report cleanly on the membrane minority. Cofactor and localization classifications add orthogonal biochemical context for construct design.

**Use as:** membrane-class confirmation and soluble-vs-membrane routing before structural modeling or construct design.

### 2.3 Residue-Level Topology

**What We Predict.** For each residue, the probability of lipid-bilayer embedding (and, from the same head, intrinsic disorder and amyloidogenicity). Evaluated on the 1,345 enzymes with UniProt transmembrane annotations — the membrane minority of the class.

**Headline Numbers.**

- Per-protein **mean AUROC 0.950**, median 0.988; AUPRC 0.802, F1 at threshold-optimal cutoff 0.798.

**Strong and Weak.** Transmembrane prediction is strong and, importantly, *uniform across EC classes*: per-class mean AUROC is 0.89–0.98 across oxidoreductases, transferases, hydrolases, lyases, isomerases, ligases and translocases alike (Figure 4). AUPRC and F1 sit below the AUROC because membrane enzymes are often single-pass — the positive (membrane) class is a small fraction of each sequence, which depresses precision-recall metrics even where the ranking is excellent. The same head supplies per-residue disorder, relevant to the flexible regulatory segments common in enzymes.

**Use as:** production triage — residue-level topology mapping for the membrane subset, ahead of construct design or structural work.

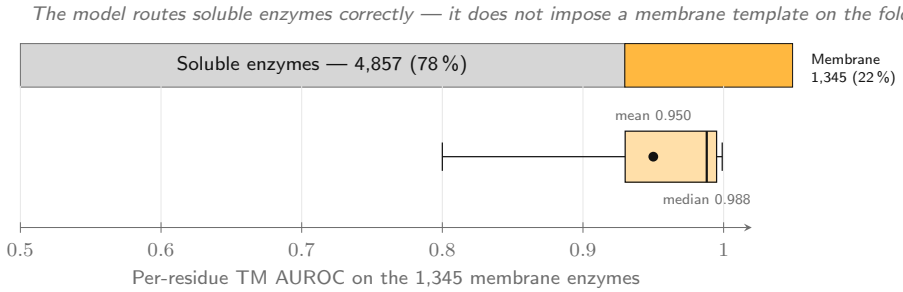



***Figure 3:*** *Topology of the enzyme class. Top: the cohort is predominantly soluble (4,857 with no transmembrane segment, 1,345 membrane). Bottom: on the membrane minority, per-residue transmembrane-segment prediction reaches mean AUROC 0.950, median 0.988.*

| | Cohort $n$ | Membrane $n$ | TM-residue AUROC |
|---|---|---|---|
| Transferases (EC 2) | *2,148* | *589* | 0.958 |
| Hydrolases (EC 3) | *1,898* | *318* | 0.956 |
| Oxidoreductases (EC 1) | *765* | *180* | 0.894 |
| Lyases (EC 4) | *209* | *32* | 0.976 |
| Isomerases (EC 5) | *206* | *18* | 0.959 |
| Ligases (EC 6) | *176* | *15* | 0.937 |
| Translocases (EC 7) | *140* | *118* | 0.965 |

Cell shade $\propto$ AUROC. Residue-level transmembrane topology is strong and uniform across all seven EC classes (0.89–0.98) — the membrane minority of each class is mapped reliably regardless of the chemistry the enzyme performs.

***Figure 4:*** *Residue-level transmembrane topology across the seven EC classes. The membrane minority of every class is mapped at AUROC 0.89–0.98, regardless of the chemistry the enzyme performs.*

### 2.4 Post-Translational Modification (PTM) Site Prediction

**What We Predict.** Per-residue probabilities for 39 PTM classes, at two operating points — a high-precision call for confident wet-lab handoff and a high-recall call 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 $\geq$1 Swiss-Prot-curated positive per modification class; enzymes provide large per-class samples (phosphorylation n=2,616, acetylation n=1,465, N-linked glycosylation n=1,142, disulfide bond n=893).

**Headline Numbers.** Figure 5 reports F1 at the threshold-optimal operating point per modification class. The high-recall operating point dominates the high-precision point on every class with sufficient ground truth — most dramatically for the structural and lipidation modifications: myristoylation F1 0.89, N-linked glycosylation 0.87, disulfide

bond 0.86.

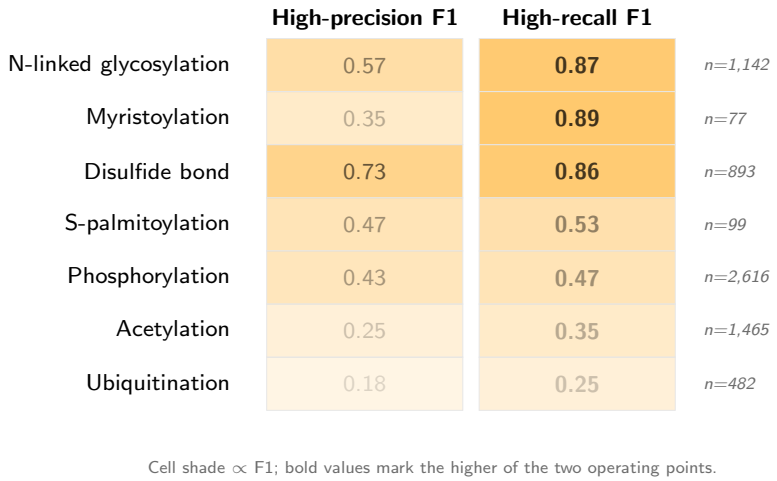



***Figure 5:*** *PTM site prediction F1 at the threshold-optimal operating point on enzyme Swiss-Prot reference, per modification class, for the high-precision and high-recall operating points.*

**Strong and Weak.** Phosphorylation is the defining enzyme modification — the largest class in the cohort by far (11,474 curated sites across 2,616 enzymes), reflecting the kinase cascades that switch catalytic activity on and off. The model ranks phosphosites at AUROC 0.95, though thresholded F1 is more modest (0.47 at the high-recall point): the conservative per-protein output budget limits recall on the most heavily phosphorylated enzymes. The structural modifications — disulfide bonds that staple secreted enzymes, N-linked glycosylation, and the myristoylation that anchors signalling enzymes to the membrane — are the strongest classes. The rare modifications (sumoylation, methylation, sulfation) sit near chance and are reported transparently (§4).

**Use as:** production triage at the high-precision operating point (wet-lab handoff); hypothesis generation at the high-recall operating point.

### 2.5 Ligand Binding Pocket Prediction

**What We Predict.** Given a target sequence, a ranked panel of plausible ligand candidates — substrate, cofactor, or inhibitor — with each ligand's predicted binding-residue set [3]. A seven-gate anti-hallucination audit verifies position validity, confidence bounds, ligand-ID realism, reproducibility, and biological plausibility on every batch. Evaluated on the 2,318 enzymes with both predictions and PDB co-crystal contacts at 4 Å.

**Headline Numbers.**

- **Ligand-identity recall: 62 %** — of PDB-observed ligand codes per protein, the model predicts 62 %.
- **Pocket-success rate 82 %**; residue F1 on ligand-ID-matched predictions 0.41, residue recall 0.57. Confidence is well-calibrated — residue recall rises monotonically with predicted confidence (0.49 at $P$ 0.70–0.85 to 0.62 at $P > 0.95$).

**Strong and Weak.** The standout result is on cofactor and nucleotide pockets — the chemically distinct, persistent active sites that define many catalytic centres. The top performers are the iron / 2-oxoglutarate dioxygenases (pirin, residue F1 1.00; PHYHD1, ALKBH6, JMJD7), the ATP pocket of a kinase (CAMK1, 0.87), the GTP pocket of a small GTPase (RAB32), and the FAD/NAD site of a flavoenzyme (CYB5R3) (Figure 6).

This is a coherent, chemistry-driven strength: a well-defined cofactor site is exactly the kind of pocket AstraBind localises best. Conversely, the broad, shallow substrate clefts of some hydrolases and the diffuse quinone sites of membrane respiratory-chain enzymes are harder, and residue precision there is lower (§4).

**Use as:** catalytic-pocket and cofactor-site localization with a ranked ligand panel; strongest for metal-, nucleotide- and dinucleotide-cofactor enzymes.

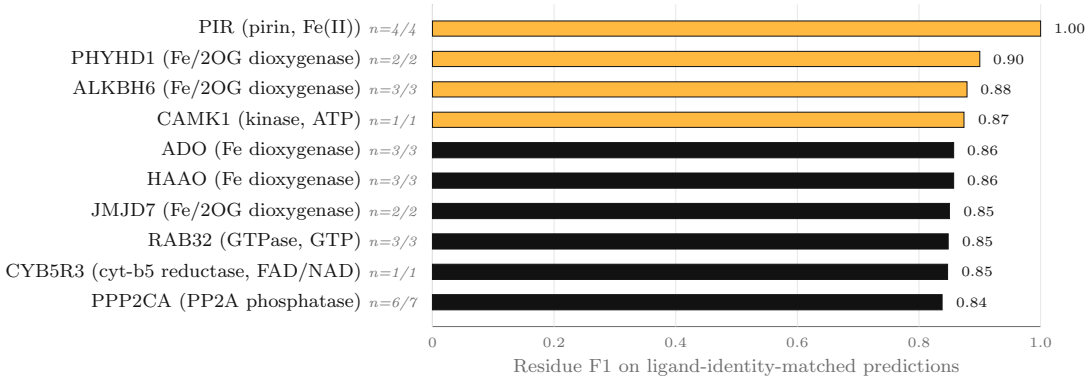



***Figure 6:*** *Top 10 enzymes by AstraBind-v2 residue F1 on ligand-identity-matched predictions. The ranking is dominated by metal-cofactor (iron dioxygenase) and nucleotide-cofactor (kinase ATP, GTPase GTP, flavoenzyme FAD/NAD) pockets, reflecting the model's cofactor-pocket strength.*

### 2.6 Thermostability Prediction ($\Delta T_m$)

**What We Predict.** For a target protein and a point mutation, the change in melting temperature ($\Delta T_m$) of the mutant relative to wild-type, in degrees Celsius, together with a confidence interval. Evaluated on **644 unique mutations across deeply-scanned enzymes**, using curated experimental thermal-shift values (FireProtDB and related stability databases) as ground truth. This is the single most valuable capability for enzyme engineering, and the suite's strongest result on this class. One scope caveat applies and we state it plainly: the benchmark proteins below are canonical, heavily-studied stability systems, so train/test overlap with public stability databases cannot be excluded for the stability model (§6); these numbers characterise performance on well-trodden targets, and the harder test is novel proteins.

**Headline Numbers.**

- On bacteriophage T4 lysozyme — the canonical protein-stability reference system, and the deepest single scan in the set (n=315) — **Spearman $\rho$ 0.93**, Pearson r 0.90, MAE 1.41 °C, sign accuracy 91 %.
- Across the full enzyme set (644 mutations, soluble and membrane enzymes): **Spearman $\rho$ 0.88**, MAE 2.64 °C, with **90 % directional accuracy** on strong-effect mutations ($|\Delta T_m| > 0.5$ °C).

**Note on the Experimental Measurement Floor.** Predictions in the small-effect band ($|\Delta T_m| < 2$ °C, either direction) are being compared against thermal-shift measurements whose own between-replicate reproducibility is comparable [12]. On enzymes with a long tail of extreme destabilisers, absolute error rises (MAE 5–6 °C) through mean-regression on the extremes even as the rank correlation stays high. For production triage, cite the rank correlation and the directional accuracy.

**Strong and Weak.** The model is most accurate on small-to-moderate effects, where T4 lysozyme is dense (e.g. the stabiliser S117V, experimental +5.1 °C, predicted +4.9; the destabiliser A98M, experimental −9.3 °C, predicted −9.4). On the extreme destabilisers it keeps the correct sign but compresses the magnitude — M102K (experimental −35 °C) is predicted at −17 °C — the documented mean-regression behaviour on extreme effects (Figure 7).

**Use as:** production pre-screen for ranking thermostabilisation mutation panels ahead of a thermal-shift assay; below actionable resolution for small effects ($|\Delta T_m| < 2\,°C$).

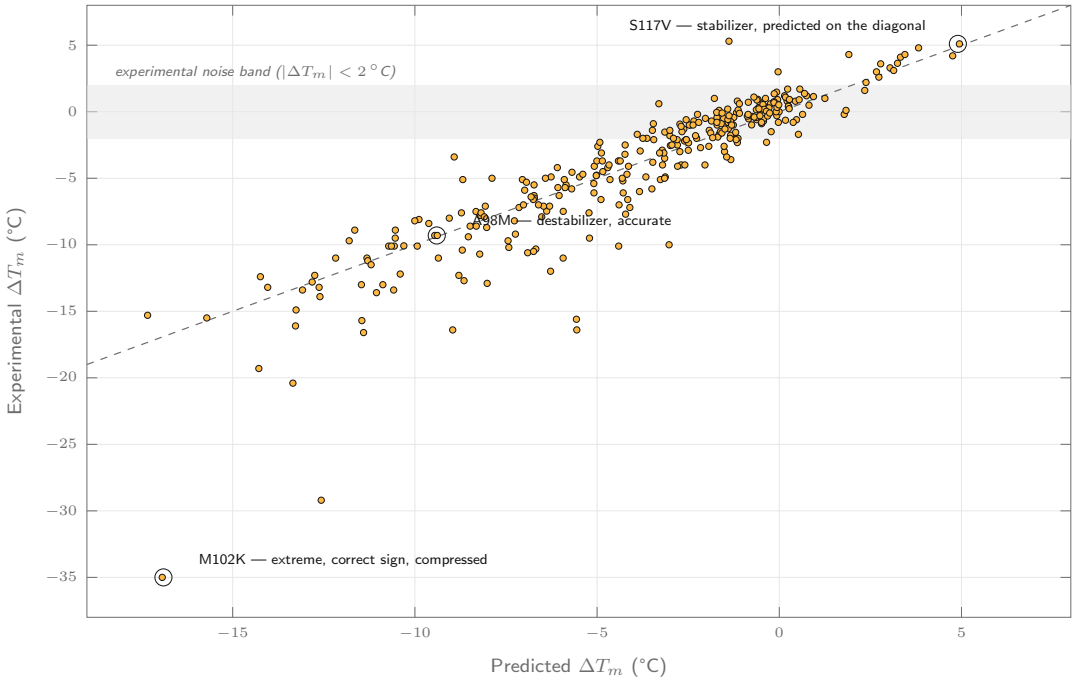



***Figure 7:*** *Predicted vs. experimentally measured $\Delta T_m$ across 315 mutations of bacteriophage T4 lysozyme (P00720), the canonical protein-stability reference system ($\rho$ 0.93). The shaded band marks the experimental measurement floor ($|\Delta T_m| < 2\,°C$); the dashed diagonal is the line of perfect prediction. A correct stabiliser, an accurate destabiliser, and one compressed extreme are annotated.*

## 3 Worked Example: A Tyrosine Kinase Across the Full Suite

The headline numbers in §2 aggregate across thousands of proteins. In practice, a discovery team works one enzyme at a time. This section walks every Astra AI model through a single, exceptionally well-characterised target — the tyrosine-protein kinase FYN (UniProt P06241, 537 residues, EC 2.7.10.2) — and shows what the integrated output looks like for a real program. FYN is a Src-family kinase that transduces signals downstream of immune and growth-factor receptors and is an oncology and neurodegeneration drug target; it carries the canonical Src architecture — an N-terminal lipid anchor, then SH3 and SH2 regulatory domains, then the bilobed catalytic kinase domain — and has a deep experimental mutational-stability dataset, which makes it an ideal end-to-end test.

Figure 8 overlays every per-area output on the kinase. The function classification returns **Transferases (EC 2) at P = 0.98** — the correct class for a protein kinase (EC 2.7.10.2) — and a top molecular-function GO term of **kinase activity**. Notably, the coarse protein-category head fires *both* **Enzymes (0.997)** and **Receptors (0.984)** — but FYN is a cytoplasmic, non-receptor Src-family kinase, so the strong "Receptors" call is

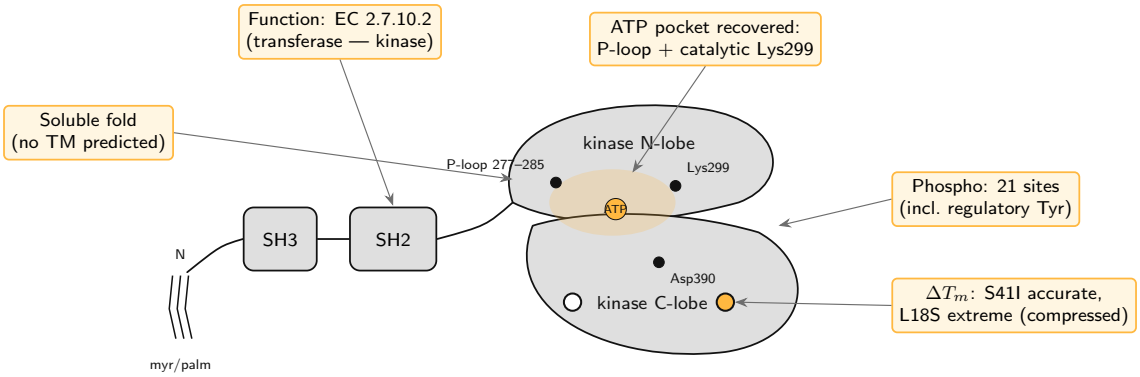



***Figure 8:*** *Integrated view of all model predictions on the tyrosine kinase FYN (P06241, 537 residues). The Src-family architecture — N-terminal myristoyl/palmitoyl anchor, SH3 and SH2 regulatory domains, bilobed kinase domain — is recovered; the predicted ATP pocket localises to the glycine-rich P-loop (277–285) and catalytic Lys299 of the inter-lobe cleft, with the catalytic Asp390 in the C-lobe; the predicted phosphorylation cluster sits on the regulatory tyrosines; two $\Delta T_m$-scan hotspots are marked. All predictions are produced from sequence alone in a single inference pass.*

category-head bleed (FYN is receptor-*associated* in signalling, not itself a receptor). This is exactly why the EC head, not the category head, is the right readout for an enzyme (§2.1).

The topology classification returns **Soluble at P = 0.97**, correctly identifying FYN as a cytoplasmic enzyme with no transmembrane segment — the model does not hallucinate a membrane span into a soluble kinase.

The PTM site prediction returns a biologically coherent set concentrated exactly where FYN is regulated and localised: a **phosphorylation cluster of 21 sites** — consistent with the kinase's regulation by autophosphorylation of its activation-loop tyrosine and by the inhibitory C-terminal tyrosine — and **N-terminal S-palmitoylation**, the lipid modification (with the co-translational myristoylation of Gly2) that anchors FYN to the membrane.

The binding-pocket prediction returns **ATP** as a top candidate ligand (confidence 0.85, 37 contact residues) and — the concrete test — its predicted contact set recovers the **entire glycine-rich P-loop (residues 277–285) and the invariant catalytic Lys299**, the ground-truth ATP-phosphate-binding machinery annotated in Swiss-Prot. This is the catalytic cofactor pocket of the enzyme, localised from sequence alone, and it is the natural starting point for an ATP-competitive inhibitor campaign.

The thermostability prediction scores FYN's deep mutational-stability scan (92 mutations) at **Spearman $\rho$ 0.90** (Pearson r 0.88). The model is accurate on small-to-moderate effects (S41I: predicted −6.9 °C, experimental −6.8; A6S: −12.7 vs −12.6; F4L: −14.5 vs −14.6) and conservative on the extremes — L18S, an experimental −43 °C destabiliser, is called with the correct sign but compressed magnitude (−19 °C), the documented mean-regression behaviour (§2.6).

**What a Program Team Would Do With This Output.** A team working FYN — or any kinase — from sequence alone can move directly to a defensible construct design and a wet-lab mutation list:

- Treat the recovered ATP pocket (P-loop + Lys299) as the orthosteric site for an ATP-competitive inhibitor screen, and the predicted contact residues as the

pharmacophore anchor.

- Preserve the N-terminal myristoylation/palmitoylation anchor if membrane localisation is required, or remove it to produce a soluble cytoplasmic construct.
- Use the predicted phosphorylation cluster to design activation-state variants (phospho-mimetic or phospho-null at the regulatory tyrosines).
- Rank thermostabilisation candidates by predicted $\Delta T_m$ before committing to a thermal-shift screen — the strong rank correlation ($\rho$ 0.90) makes the prediction a reliable prioritiser, keeping mutations clear of the catalytic and ATP-binding residues identified above.

The output yields a ∼20–40 variant panel and a construct design ready for wet-lab follow-up, starting from nothing but the sequence.

## 4 Limits and Operating Envelope

Four operating boundaries are worth making explicit for the enzyme class. Each is a scope statement, not a recovery roadmap.

**Read the EC Head, Not the Category Head.** The suite carries two routes to "is this an enzyme": a coarse protein-category head and the fine-grained EC head. They disagree, and the disagreement is informative. The category head labels 68.8 % of the cohort "Enzymes" — not because it misses the rest, but because a large fraction of enzymes carry a *dominant* non-catalytic identity: a receptor tyrosine kinase is first a signalling protein, a regulatory protease is first a regulatory protein, a structural enzyme is first structural. The EC head, which asks the enzyme-specific question directly, recognises 95.3 % as catalytic — only 4.7 % are called non-enzyme — with a mean $P$(enzyme) of 0.95. For an enzyme program, read the EC head; the category head is answering a different, broader question and will under-count by design.

**Binding-Pocket Prediction Is Strongest for Well-Defined Cofactor Pockets.** Ligand-identity recall is carried by chemically-distinct, persistent pockets — above all the nucleotide and dinucleotide cofactor sites (the ATP pocket of the kinases, the NAD(P)/FAD sites of the oxidoreductases). The harder case is the broad, shallow substrate clefts of some hydrolases and the catalytic surfaces that are only completed upon substrate binding; here residue precision is lower. The prediction localises the cofactor and catalytic machinery well; it is not a substitute for a co-crystal structure of a specific inhibitor (§2.5).

**Thermostability Operating Envelope — The Experimental Measurement Floor.** The flagship rank correlation ($\rho$ up to 0.93) rests on the most deeply-scanned enzyme in the set, and the strong-effect directional accuracy is excellent. Two honest caveats apply. Predictions in the small-effect band ($|\Delta T_m| < 2\,^\circ$C, either direction) are being compared against a thermal-shift measurement whose own between-replicate resolution is comparable — a property of the evaluation, not the model. And on enzymes with a long tail of extreme destabilising mutations (the kinase scans), absolute error rises (MAE 5–6 °C) through mean-regression on the extreme effects even as the rank correlation stays high ($\rho$ 0.88–0.91). For production triage, cite the rank correlation and the directional accuracy, not the absolute error on outliers.

**PTM Coverage Is Uneven Across the 39 Classes.** The modification classes with abundant Swiss-Prot ground truth in enzymes — phosphorylation, disulfide bonds, N-linked glycosylation — are predicted well (§2.4). The rare-modification classes (sumoylation, methylation, sulfation) sit near chance on the high-precision operating point and

are reported transparently rather than suppressed. The suite is built to *miss rather than misflag* on these classes: a low-recall, high-precision default that returns few false positives is the right behaviour when the downstream cost is an experimental follow-up. Choose the high-recall operating point only when exhaustive candidate enumeration is the goal.

## 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 enzyme program questions to the prediction-area combinations that address them, the operating points to use, and the expected output for a discovery or protein-engineering team. Numbers in parentheses reference the per-area evidence in §2.

***Table 1:*** *Decision framework mapping common enzyme 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. Engineering an Enzyme for Thermostability** | This is the flagship use case for the class. The **thermostability prediction** ranks candidate point mutations by predicted $\Delta T_m$ (rank correlation $\rho$ up to 0.93 on the deepest single-enzyme benchmark, §2.6); prioritise the top-ranked stabilisers for a thermal-shift screen. Cross-check positions against the **ligand/active-site pocket prediction** to keep mutations away from catalytic and cofactor-binding residues, and against the **residue-level fold prediction** to favour ordered core positions over disordered regions. A 20–40 variant panel for thermal-shift validation is the typical output. |
| **2. Active-Site, Cofactor & Inhibitor Pocket Identification** | The **ligand binding pocket prediction** returns a ranked candidate-ligand panel with predicted contact residues, localising the catalytic cleft, cofactor-binding motif, and druggable inhibitor pocket. Cross-check the predicted pocket against the **function classification** (the EC class constrains the expected chemistry and cofactor) and against the **residue-level fold prediction** to confirm the site sits in an ordered domain. The output scopes the catalytic machinery before a mechanistic or medicinal-chemistry campaign. |
| **3. Function Assignment for an Uncharacterised Enzyme** | The **function and family classification** assigns the EC top-class (EC 1–7; 95.3 % of the cohort recognised as enzymatic) and the molecular-function GO terms (top-5 ancestor accuracy 94.2 %). Combine with the **ligand-pocket prediction** to corroborate the predicted chemistry against a physical cofactor/substrate site, and with **PTM site prediction** to flag regulatory modifications. This scopes a newly sequenced or hypothetical enzyme before wet-lab assignment. |

| Program Question | Recommended Workflow Across Orbion's Astra AI Suite |
|---|---|
| **4. Enzyme Disease-Variant Interpretation** | Many enzymes are Mendelian disease genes (the inborn errors of metabolism). The **thermostability prediction** estimates the fold-stability impact of each missense variant — the dominant loss-of-function mechanism for enzymes is destabilisation of the fold rather than direct active-site disruption; combine with the **ligand/active-site pocket prediction** to separately flag variants that hit catalytic residues, and **PTM site prediction** to catch variants that abolish regulatory sites. ClinVar or gnomAD variants can be re-ranked by a combined Astra-derived deleteriousness score. |
| **5. Mapping Enzyme Regulation (PTM Landscape)** | Enzymes are regulated heavily by phosphorylation — the largest modification category in this class (§2.4). The **PTM site prediction** (high-recall operating point) returns the candidate-site set across all 39 modification classes; focus on the phosphorylation that switches catalytic activity on and off and the disulfide bonds that staple secreted enzymes. Intersect with the literature-curated set in Swiss-Prot to surface novel candidate sites for mass-spectrometry follow-up. |

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. 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 $\Delta T_m$ 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.** 6,304 reviewed enzyme-associated proteins from UniProt Swiss-Prot [4], drawn from the catalytic-activity / enzyme annotation across all organisms, de-duplicated against

sequence-integrity checks. Sequences with non-standard residues (X/B/Z/U/O) or outside the 16–5000-residue length window were excluded at preflight. EC top-class assignments follow the Enzyme Commission's seven-class hierarchy; functional groupings in §2.1 are by the model's predicted EC class.

**Experimental Reference.** UniProt feature table for sequence-level features (active sites, binding sites, modified residues, glycosylation sites, disulfide bonds, transmembrane segments, regions). Gene Ontology molecular-function annotations expanded to the is_a / part_of ancestor closure via QuickGO [6, 7]. PDB ligand-contact reference (§2.5): heavy-atom contacts at 4.0 Å between each ligand — substrate, cofactor, or inhibitor — and protein residue, aggregated across all deposited co-crystal structures per UniProt accession [5]. $\Delta T_m$ reference (§2.6): curated experimental thermal-shift and melting-temperature mutation scans from FireProtDB and related public stability databases, spanning soluble and membrane enzymes — the bacteriophage T4 lysozyme stability dataset (the canonical reference system for protein-stability prediction) together with ribonuclease, protein-kinase, and rhomboid-protease scans. De-duplicated to one prediction per (protein, mutation) pair, yielding 644 unique mutations.

**Metrics.** AUROC, AUPRC [13], F1 at threshold-optimal cutoff, 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. The 5 GB per-protein prediction cache was processed by streaming (one protein at a time) to keep memory bounded.

**Evaluation Scope.** The classification, topology, and PTM metrics in §2.1–§2.4 measure how the deployed Astra models perform on the enzyme class as they are run in production. The cohort is drawn from the reviewed proteome, which overlaps the corpora these models were trained on; the figures therefore characterise within-distribution behaviour on real targets rather than held-out generalisation. Held-out test-split performance for each model is reported in its respective preprint [1–3]. The contribution of this report is the per-EC-class enzyme performance breakdown and the integrated single-protein workflow (§3). The thermostability evaluation (§2.6) is reported against curated experimental thermal-shift measurements rather than annotation recall. We do *not* claim it is held-out: the deeply-scanned benchmark proteins (T4 lysozyme, ribonucleases, Src-family kinase) are canonical stability systems that are well-represented in the public mutation databases used to train stability predictors, so we cannot certify the model's training data is disjoint from these proteins. The $\Delta T_m$ figures therefore characterise performance on well-studied targets; a genuinely held-out estimate requires novel proteins, and is the priority of ongoing validation.

**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] 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] Matthews, B. W. Studies on protein stability with T4 lysozyme. *Advances in Protein Chemistry*, **46**:249–278 (1995). doi:10.1016/S0065-3233(08)60337-X.

[9] Stourac, J., Dubrava, J., Musil, M., Horackova, J., Damborsky, J., Mazurenko, S., Bednar, D. FireProtDB: database of manually curated protein stability data. *Nucleic Acids Research*, **49**(D1):D319–D324 (2021). doi:10.1093/nar/gkaa981.

[10] Nikam, R. *et al.* ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Research*, **49**(D1):D420–D424 (2021). doi:10.1093/nar/gkaa1035.

[11] McDonald, A. G., Tipton, K. F. Enzyme nomenclature and classification: the state of the art. *FEBS Journal*, **290**(9):2214–2231 (2023). doi:10.1111/febs.16274.

[12] Niesen, F. H., Berglund, H., Vedadi, M. The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature Protocols*, **2**(9):2212–2221 (2007). doi:10.1038/nprot.2007.321.

[13] 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.