A003 · Elucidating enzyme–substrate specificity through co-folding foundation model

Presenter: Xiwei Cheng — Northeastern University

Authors: Xiwei Cheng, Seonghwan Seo, Charlie Huh, Jihang Chen, Songlin Jiang, Pengkang Guo, Jing-Ke Weng, Woo Youn Kim, Wengong Jin

Session: Day 2 · Fri Oct 2, 2026 · 2:15–4:15 PM

Enzymatic catalysis relies on precise structural and chemical complementarity, yet systematically mapping enzyme-substrate interactions remains a critical bottleneck. While structure-aware methods have advanced functional annotation, their reliance on predefined binding pockets and rigid-body docking fails to capture the ligand-induced conformational changes essential for catalytic turnover. Here we introduce Boltz2ESI, an end-to-end framework that predicts enzyme–substrate interactions by leveraging structural knowledge learned by a biomolecular foundation model. Through native co-folding, the framework inherently captures active-site plasticity without requiring predefined pocket annotations. Integrating these learned biophysical priors with global evolutionary context and geometric molecular descriptors, Boltz2ESI consistently outperforms state-of-the-art sequence-based and rigid-docking approaches. Extensive validation demonstrates that the framework accurately discriminates tight subfamily specificities, enabling effective candidate prioritization for biosynthetic pathway elucidation, as demonstrated on the withanolide pathway. Ultimately, this structure-dynamic approach establishes an actionable foundation for accelerating rational biocatalyst discovery and large-scale pathway de-orphaning.

Elucidating enzyme–substrate specificity through co-folding foundation model

Xiwei Cheng[1,2,*,†] Seonghwan Seo[3,†] Charlie Huh[4,5] Jihang Chen[1] Songlin Jiang[1] Pengkang Guo[6] Jing-Ke Weng[4,5] Woo Youn Kim[3,7,8] Wengong Jin[1,2,4]

[1]Khoury College of Computer Sciences, [4]Institute for Plant–Human Interface, [5]Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA; [2]Broad Institute of MIT and Harvard, Cambridge, MA, USA; [3]Department of Chemistry, [8]InnoCORE AI-CRED Institute, KAIST, Daejeon, Republic of Korea; [6]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland; [7]HITS Inc., Seoul, Republic of Korea

[*]Presenting author: cheng.xiw@northeastern.edu [†]These authors contributed equally Correspondence: w.jin@northeastern.edu

Introduction. Enzymatic catalysis underpins metabolism and industrial biocatalysis, yet mapping which substrates an enzyme accepts remains a bottleneck for biocatalyst discovery. Sequence-based predictors cannot see the catalytic microenvironment, while structure-aware methods rely on predefined pockets and rigid docking: they fail for orphan enzymes with uncatalogued active sites and cannot represent the ligand-induced plasticity that dictates selectivity. We present Boltz2ESI, which resolves specificity directly from co-folded enzyme–substrate complexes generated by a biomolecular foundation model, capturing active-site plasticity without predefined pocket annotations and reaching state-of-the-art accuracy on ESIBank.

Method. Given only an enzyme sequence and a substrate SMILES string (Fig. 1), two-stage co-folding first localizes the active site with Boltz-2 [1], then re-folds the cropped pocket together with the substrate in an alignment-free regime, so the pose is resolved from mutual biophysical complementarity rather than evolutionary sequence bias. Single, pair, and distogram representations of the co-folded complex are fused with ESM3 [2] evolutionary embeddings and Uni-Mol2 and Morgan substrate priors, then passed through a geometry-conditioned four-block Pairformer initialized from the Boltz-2 affinity head. Training uses ESIBank [3] (323,783 curated pairs).

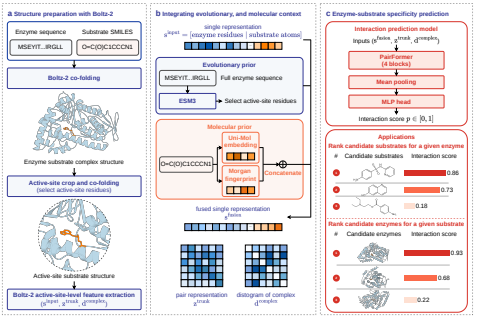

Figure 1: Boltz2ESI workflow. Two-stage Boltz-2 co-folding localizes and re-folds the active site and extracts structural representations (**a**); evolutionary and molecular priors are fused into a unified representation (**b**); a geometry-conditioned Pairformer outputs the interaction score that ranks candidate substrates or enzymes (**c**).

Results. Boltz2ESI surpasses all baselines on both ESIBank splits (Fig. 2a,b): AUROC 0.9156 / AUPR 0.6221 (random split) and AUROC 0.7664 (unknown substrate and enzyme) versus 0.7198 (EZSpecificity [3]) and 0.6778 (ESP [4]). Zero-shot Boltz-2 affinity is near random (0.5258 / 0.5299), indicating that affinity alone does not explain catalytic specificity. Ablations (Fig. 2c) show evolutionary context and co-folded geometry are complementary: removing ESM3 causes the largest AUROC drop, while replacing co-folded complexes with rigid docking poses collapses AUPR from 0.3331 to 0.2169.

Pathway reconstruction in Withanolide biosynthesis. Withanolides are the phytosteroidal lactones that constitute the primary bioactive metabolites of *Withania somnifera* (ashwagandha), a widely used dietary supplement, and are reported to have anti-inflammatory, neuroprotective, and anticancer activity. Their biosynthesis was elucidated only recently, when three cytochrome P450s, CYP87G1, CYP88C7, and CYP749B2, were shown to catalyse the key early oxidations [6]. Identifying which member of a large transcriptome-derived P450 family performs a given step is exactly the prioritization problem Boltz2ESI addresses. Without pathway-specific training, Boltz2ESI ranked CYP87G1 8th and CYP88C7 4th among 107 candidates from a *de novo W. somnifera* transcriptome, ahead of EnzymeCAGE [5] (13th and 10th; Fig. 2d), recovering every biochemically characterized paralog of both enzymes within the top 20 and top 10, respectively. Because it scores structural complementarity rather than co-expression, Boltz2ESI complements omics-based pathway inference for de-orphaning enzymes and prioritizing biocatalysts.

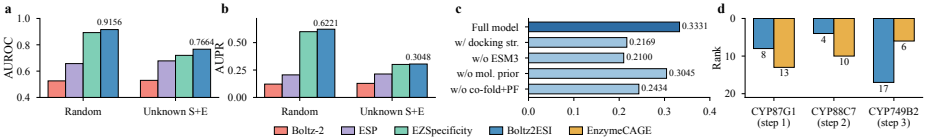

Figure 2: Benchmark, ablation, and pathway prioritization. a,b, AUROC and AUPR on the ESIBank random and unknown-substrate-and-enzyme (S+E) splits. **c,** Ablation on the unknown S+E split (AUPR; PF, Pairformer). **d,** Rank of each validated CYP among 107 *W. somnifera* P450 candidates (lower is better).

References. [1] Passaro *et al. bioRxiv* 2025. [2] Hayes *et al. Science* 2025. [3] Cui *et al. Nature* 2025. [4] Kroll *et al. Nat. Commun.* 2023. [5] Liu *et al. bioRxiv* 2024. [6] Reynolds *et al. Nat. Plants* 2026.