DeepAden: an explainable machine learning for substrate specificity prediction in nonribosomal peptide synthetases

Jiaquan Huang[1#], Liangjun Ge[1#], Yaxin Wu[1#], Qiandi Gao, Pan Li, Jun Wu, Heqian Zhang[1*], Zhiwei Qin[1*]

[1]Center for Biological Science and Technology, Advanced Institute of Natural Sciences, Beijing Normal University, Zhuhai, Guangdong, 519087, China.

[#]These authors contributed equally

*Correspondence: z.qin@bnu.edu.cn; zhangheqian@bnu.edu.cn

Abstract

Microbial nonribosomal peptide synthetases (NRPSs) assemble diverse peptides via adenylation (A) domains that activate various substrates. While existing tools predicting substrate specificity of A-domain incorporate machine learning methods and specificity-conferring codes, they inadequately address substrate promiscuity and non-proteinogenic substrates. Recent advances in cross-modal contrastive learning combined with few-shot learning offer promising approaches to improved prediction of these complex pocket-substrate interactions. Here we present DeepAden, a novel two-stage ensemble tool that achieves state-of-the-art performance in substrate specificity prediction across multiple benchmarks. The framework first trains a graph attention network to identify A-domain binding pocket around substrate using fused sequence and structural features derived from a protein language model, then employs cross-modal contrastive learning to jointly encode pocket-substrate interactions through protein and chemical language model, effectively addressing both substrate promiscuity and zero-shot prediction challenges. Furthermore, we propose an explainable artificial intelligence (XAI) strategy to serve dual purposes, which not only guides position-specific substitutions for data augmentation, but also enables systematic investigation of substrate-specific residues across characterized A-domains. XAI analysis revealed Cys-ABPs uniquely employ ^{18}Gly and ^{20}Ala, in contract to ^{18}Tyr/^{20}Pro in other ABPs, forming a flexible binding pocket that enables distinctive ligand orientation through altered hydrogen bonding networks and π-π stacking interactions. Collectively, our work establishes an XAI-powered paradigm that resolves A-domain substrate specificity at residue-level precision, transforming black-box predictions into actionable blueprints for future rational NRPS genome mining and engineering.

Introduction

Nonribosomal peptides are naturally occurring biopolymers exhibiting remarkable structural and functional diversity, serving roles as antibiotics, siderophores, toxins and immunosuppressants[1]. The assembly of their peptide backbones is mediated by highly modular megasynthetases, termed nonribosomal peptide synthetases (NRPSs), which utilize amino acids as monomers (commonly referred

to as building blocks)[2]. NRPSs are minimally composed of three core domains: adenylation (A) domain responsible for substrate selection and activation, eptidyl carrier protein (PCP) domain that tethers activated amino acids via its phosphopantetheinyl prosthetic group, and condensation (C) domain that catalyzes peptide bond formation to elongate the chain[3]. In principle, the sequential arrangement of A-domains within biosynthetic gene clusters (BGCs) determines the peptide sequence; thus, understanding their specificities is critical for predicting peptide structures[4]. However, studying A-domain in isolation encounters significant experimental and computational challenges. To address this, researchers have developed specificity-conferring codes (SCCs) and A-domain binding pocket (ABP) residues, including those targeting first-shell residues (direct substrate recognition)[5] and second-shell residues (binding pocket stabilization)[6], and combined these residues with computational approaches such as sequence alignment, hidden Markov models (HMMs)[7], and support vector machines (SVMs)[8] to predict substrate specificity. Despite these efforts, the development of NRPS A-domain prediction algorithms face three key challenges: (1) SCCs or ABP residues derived from co-crystallization analyses of A-domain in a single species may lack generalizability across divergent species; (2) the sequence diversity of A-domains limits the accuracy of alignment-dependent methods; and (3) existing machine learning models struggle to handle datasets with nonproteinogenic substrates and substrate promiscuity. Therefore, addressing these challenges is crucial for advancing NRPS biosynthesis research and enabling the engineering of novel peptide-based therapeutics.

Existing A-domain prediction tools, such as NRPSpredictor2[9], SANDPUMA[10] and AdenPredictor[11], can effectively identify proteinogenic substrates, but are limited in identifying non-proteinogenic substrates, probably because they rely on sequence alignment based on the SCCs template of GrsA-Phe[12]. These methods, which require high sequence similarity between target and reference domains, often fail to predict SCCs and ABP residues in evolutionarily divergent sequences, highlighting the necessity of incorporating structural information rather than relying on primary sequence alignment[13]. To advance substrate specificity prediction, it is essential to develop robust methods to identify ABP residues as an alternative to traditional SCCs, enabling more accurate prediction across diverse A-domains regardless of evolutionary distance. Deep learning algorithms have revolutionized computational biology by enabling high-accuracy analysis of complex biological data[14]. Protein language models (PLMs) and chemical language models (CLMs), such as ESM-2 and MoLFormer, leverage self-supervised learning to uncover evolutionary patterns from vast protein and molecule datasets[15]. Transfer learning techniques further allow the extraction of structural and functional features from PLMs, facilitating applications like residue mutation effect and protein function prediction[14c, 16]. For example, utilizing the ESM2-650M model achieves accurate prediction of residue-residue contacts in both active and inactive peptides[17], which directly facilitates the evolution of antimicrobial peptides. Although these AI models bypass sequence alignment limitations, their lack of interpretability often termed the “black-box” problem, hinders their utility in guiding mechanistic insights, such as identifying binding pocket residues critical for substrate recognition[18]. Recent

explainable artificial intelligence (XAI) methods like SHAP (SHapley Additive exPlanations) have emerged to decode model decisions by quantifying the contribution of individual amino acids to protein properties[19]. Such methods could potentially address the substrate specificity of evolutionarily divergent proteins, a challenge that has long confused natural product chemists.

Current technical limitations in substrate specificity prediction of A-domain primarily stem from conventional algorithms such as SVMs and tree-based models, which heavily depend on training data quality and quantity[20]. These limitations become more evident when predicting nonproteinogenic substrates, where limited experimental data prevents effective learning of structural features, creating gaps in prediction coverage. Notably, the chemical structural novelty of nonproteinogenic substrates is a critical feature of novel NRPs. These non-proteinogenic substrates serve as essential components of bioactive oligopeptides, demonstrating significant potential in protein engineering and drug discovery[12, 21]. Improving model performance under few sample conditions, called few-shot learning, would enhance prediction generalizability and accelerate the discovery of novel NRPS-derived natural products. Recent advances in pretrained PLMs and CLMs have transformed the study of ligand-protein interactions[22]. Emerging alignment-free approaches, inspired by virtual screening, directly predict protein-substrate interactions using molecular descriptors (e.g. SMILES strings) and protein sequences, eliminating the need for 3D structural data. State-of-the-art models trained on binding affinity data can systematically screen optimal interaction pairs, while contrastive learning frameworks integrating PLMs with CLMs potentially enhance robustness under data scarcity[23]. By projecting protein and molecular features into a unified vector space, these frameworks might enable rapid pairing of A-domains with compatible substrates.

To address the limitation mentioned above, we present DeepAden, an two-stage ensemble tool for A-domains prediction, comprising three innovative modules: (1) a ResNet-based residue contact predictor that leverages spatial constraints from protein language model attention maps; (2) A graph attention neural network-based A-domain binding pocket prediction model (ABP-GAT) that integrates residue features with probabilistic contact edges to incorporate dynamic binding pocket changes for identifying template-free ABPs; and (3) a contrastive learning framework, enabling cross-modal representation alignment of ABPs and molecular fingerprints while addressing dataset imbalance through with XAI-guided data augmentation. Benchmark evaluations demonstrate that DeepAden quantifies substrate promiscuity through similarity pairing and achieves exceptional performance on nonproteinogenic substrates. Compared to conventional tools like NRPSpredictor2, DeepAden improves prediction accuracy and applicability for substrate specificity characterization. With the assistance of XAI and structural analysis, while most ABPs contain ^{18}Tyr/^{20}Pro, Cys-ABPs instead employ ^{18}Gly and ^{20}Ala confers flexibility to the binding pocket, allowing an alternative ligand orientation due to altered hydrogen bonding and π–π stacking interactions. Our work establishes DeepAden as a reliable tool for A-domains specificity prediction, overcoming limitations of alignment-dependent methods and laying

the foundation for discovering novel NRPs and their biosynthetic logic.

Results

DeepAden overview

To achieve both generalizability and specificity in predicting substrates of A-domains, DeepAden integrates the protein language models (PLM) ESM2 with a graph attention neural network (GAT) to accurately identify binding pockets within the A-domain, even without structural information. Specifically, DeepAden constructs a protein graph directly from the input protein sequence, embedding graph nodes by combining residue embeddings derived from ESM2 with physicochemical properties. Graph edges are defined based on a contact map predicted by a ResNet model trained on the ESM2 attention map. Binding pocket regions are annotated across the entire graph, and the GAT is employed to predict the nodes corresponding to binding pocket residues. Consistency-based semi-supervised techniques are incorporated into the model framework to mitigate overfitting in node prediction (Fig. 1A). DeepAden reformulates substrate prediction as a dense retrieval task that computes a contrastive loss between two separate pre-trained encoders: ABP-ESM2 for binding pockets sequence and MoLFormer for substrate' s simplified molecular-input line-entry system (SMILES) string. This approach maximizes the similarity between a pocket–substrate pair if the pocket adenylates the substrate and minimizes it otherwise (Fig. 1A). Moreover, we fitted a two-component Gaussian Mixture Model (GMM) on the scaled cosine similarities between pocket embeddings and molecule embeddings to quantify the confidence of the prediction results and provide flexible selection (Fig. 1B).

DeepAden achieves SOTA on binding pocket prediction using graph attention neural network

Traditional discrete coding approaches like the 10-AA[5] and its derived 13-AA[24], 17-AA[25] SCCs face limitations in predicting diverse substrates due to evolutionary divergence in A-domain structures. The 34-AA SCCs represented a significant advancement by incorporating structural elements within an 8Å radius of substrate in GrsA-Phe (PDB: 1amu), establishing its current implementation in tools such as NRPSpredictor2 embedded in antiSMASH v7.0. However, existing prediction methodologies not only rely on limited training datasets primarily focused on proteinogenic amino acids but also face inherent challenges in applying HMM algorithms to non-proteinogenic substrates due to the absence of appropriate templates. To address these limitations, we conducted a comprehensive structural analysis of 13 co-crystal complexes containing both proteinogenic and non-proteinogenic substrates (Table S1). Through systematic evaluation of residues within a 6Å radius of the binding pocket, we identified a discrete set of 27 ABP residues, distributed in three β-sheets, two α-helices, and a loop connecting a β-sheet and an α-helix (Fig. S1). As demonstrated in Fig.2A, these six regions collectively form the cavity structure of the binding pocket.

To establish a comprehensive repository of A-domain with substrate information, we curated 786 experimentally validated A-domains from MIBIG 3.1, encompassing 115 types of substrates including

96 non-proteinogenic substrates (Table S2 and Fig. S2). Current computational tools, including NRPSpredictor2 and AdenPredictor, have not adequately addressed these non-proteinogenic substrates categories, primarily because their SCCs coding was challenging to be extracted using HMM algorithms, and available training data remained limited. To overcome the limitations of HMM algorithms, which exclusively rely on primary sequence alignments for binding pocket predictions, we utilized sequence and structural information from 78 published crystal structures to developed a ResNet-based contact map prediction model (F1 score = 0.965 in test dataset) to capture residue-level interactions within the A-domains (Fig. 2B). With this structural prediction capability in sequence level, we incorporated node features provided by PLM and edge features provided by our contact map prediction model to develop ABP-GAT (F1 score = 0.979 in test dataset). In this study, we investigated all 786 A-domains exhibiting 20% to 60% identity to the 1amu sequence to compare the binding pocket site prediction accuracy between the HMM algorithm and the ABP-GAT model. Results demonstrated that as identity to the 1amU sequence decreased, the ABP residues prediction accuracy of the HMM algorithm declined significantly, while the ABP-GAT model maintained highest prediction accuracy across all ranges (Fig. 2C). This performance stability was particularly pronounced when identifying binding pocket for non-proteinogenic substrates, where traditional methods typically struggle the most.

Conservative and functional residues recognition through evolutionary and explainable AI analysis

Employing the ABP-GAT model, we efficiently generated a dataset of 27-AA ABP sequences, comprising 786 labeled sequences with substrate information and 186,758 unlabeled sequences without substrate information, using exclusively sequence-based inputs. To evaluate the functional relevance of these ABP sequences in substrate pairing, sequence similarity networks (SSNs) were constructed using a 70% sequence identity threshold (Fig. 3A). These SSNs revealed phylogenetically distinct clusters corresponding to specific substrate classes, including β-amino acids, aromatic acids, α-keto acids, and carboxylic acids. Notably, ABPs associated with hydrophilic amino acids (e.g. Ser, Cys, Thr, Gly) and non-proteinogenic amino acids (e.g. dHpg, Hdpg) formed discrete, substrate-specific clusters. In contrast, hydrophobic amino acids (e.g. Val, Ile, Leu) and acidic amino acids (e.g. Asn, Asp) were associated with substrate promiscuity, suggesting broader binding adaptability. Comparative analysis of amino acid conservation patterns between the labeled and unlabeled datasets identified key structural determinants. Specifically, positions 2, 3, 15, 19, and 22 exhibited high conservation in labeled pockets, while positions 2, 3, 7, 20, 24, and 25 showed potential evolutionary divergence between labeled and unlabeled A-domains (Fig. 3B and 4A). These findings confirm that ABP sequences encode sufficient information to distinguish substrate preferences. To advance mechanistic understanding, it is imperative to systematically characterize how positional residues contribute to substrate specificity. Such analyses could elucidate the evolutionary drivers of substrate selectivity and promiscuity across enzyme families.

Recent studies have leveraged Long Short-Term Memory (LSTM) networks for peptide embedding and SHapley Additive exPlanations (SHAP) to quantify the contributions of individual amino acid residues

to model predictions, enabling the identification of functionally critical sites governing peptide properties. Building on this approach, we fine-tuned the ESM2-650M model, referred to as ABP-ESM2, using a curated dataset of 187,544 ABP sequences. This model generates context-aware positional embeddings, which demonstrate enhanced discriminative capacity in embedding distributions compared to the original ESM2 model (Fig. S3), effectively preserving positional information critical for downstream analyses. For the labeled set of 786 ABP sequences, we implemented an oversampling strategy combined with 5-fold cross-validation to develop ABP-RF, a robust random forest classifier that achieved an accuracy of 0.929. It is noteworthy that ABP-RF employs SHAP analysis without positional indexing of input residues, as the ABP-ESM2 embeddings inherently retain contextual information. This framework allows direct attribution of predictive contributions to specific residue positions. Overall SHAP analysis across all classes revealed that the ten most contributive residues (positions 20, 25, 2, 15, 21, 5, 22, 18, 24, and 3), located proximal to non-carboxyl substrate moieties, side-chain interaction zones, and AMP regions, exerted the greatest influence on substrate specificity (Fig. 4A and 4B). Notably, all these residues are part of the 34-AA SCCs defined by Rausch et al. and extend the canonical 10-AA SCCs defined by Stachelhaus et al (Table S3). Of particular interest, residues at positions 15 and 18, located near the AMP-binding interface, suggest that regions beyond the classical active site may modulate substrate-ATP spatial coordination, therefore potentially influence substrate specificity of A-domains.

Substrate-specific SHAP value analysis elucidated the molecular mechanism underlying substrate specificity recognition in A-domains. For clarity in the following description, the superscript numbers preceding the amino acid residues indicate their positions within the ABP sequence alignment, while the numbers following the residues correspond to their positions in the resolved PDB structure. For example, L-Cys and L-Ser have only minor differences in side chain size and polarity, therefore the mechanism by which A-domains distinguish between these two amino acids is not well understood. To address this issue, the comparative analysis of SHAP value distributions between Cys-ABPs and non-Cys-ABPs indicate that differences in substrate binding site composition and changes in AMP spatial orientation are key factors in Cys recognition. A systematic evaluation of SHAP values and amino acid substitutions among Cys-ABPs and other amino acid substrates-ABPs (Ser, Val, Gly, Pro, Ala, Thr, His, Tyr, Asn, Gln) demonstrated conservation of ^{18}Gly and ^{20}Ala in Cys-ABPs, contrasting with ^{18}Tyr and ^{20}Pro in other substrate-ABPs (Fig 4C and Fig. S4). These residues emerge as critical determinants of substrate specificity. High-resolution structural comparisons between co-crystallized complexes 7en1 (Cys-AMP-bound) and 6ea3 (Ser-AMP-bound) uncovered distinct ligand coordination mechanisms (Fig 5A and 5C). In the Ser-ABP structure, the adenine ring of AMP engages in stable π-π stacking interactions with ^{18}Tyr, while its N3 atom and amino group form hydrogen bonds with ^{15}Gly and ^{17}Leu, respectively (Fig 5C and 5D). These interactions position the adenine ring nearly parallel to the ^{15}Gly-containing loop. Concurrently, ^{20}Pro imposes steric hindrance on the ribose-phosphate moiety, restricting its flexibility (Fig 5C). In contrast, the Cys-ABP adopts a distinct configuration, where the adenine ring engages in an "edge-to-face" π-π stacking interaction with Phe939, located outside the binding pocket (Fig 5A and 5B).

This interaction tilts the adenine ring upward, enabling two hydrogen bonds to form with ^{17}Leu (Fig 5B). Consequently, the ribose moiety undergoes a structural reorientation, with its O4' atom repositioned closer to the adenine ring, in contrast to its configuration in the Ser-ABP (Fig 5A). Notably, ^{15}Gly in Cys-ABP no longer directly interacts with the adenine ring but instead mediates the conformational coupling between the ribose and the phosphate monoester (Fig 5A and 5B). Additionally, the presence of ^{20}Ala and ^{18}Gly in Cys-ABPs imparts greater local flexibility to the adjacent α-helix and loop, compared to the rigid ^{20}Pro and bulky ^{18}Tyr in Ser-ABP (Fig 5A and 5C). This enhanced flexibility allows the thiol (-SH) group of Cys to approach ^{19}Gly, forming a stable hydrogen bond (Fig 5A). These structural insights elucidate the precise discrimination between Cys and Ser substrates with similar side-chain volumes and polarities. Furthermore, this mechanism explains the distinct clustering of Cys-ABPs in sequence similarity network (SSN) analyses, reflecting evolutionary divergence from other substrate-ABPs.

DeepAden achieves SOTA in single Substrate prediction through XAI-guided data augmentation and contrastive learning

Current computational approaches for predicting substrates of NRPS A-domains exhibit unsatisfactory performance for recently identified non-proteinogenic substrates. This limitation derives not only from inaccurate substrate SCC extraction methods but more critically from the extreme scarcity of experimental data for these substrates. Traditional data augmentation techniques in biological contexts, such as random perturbations, unpredictably destabilizing molecular or protein structures, often yielding chemically implausible data. In this study, we identified evolutionarily conserved and functionally critical residues that drive class-specific predictions in individual samples. Notably, residues that do not constitute defining features of their assigned classes present viable targets for conservative mutagenesis simulations. Such targeted mutagenesis simulations enable dataset expansion while preserving structural and functional rationality. To operationalize this strategy, residues exhibiting the three lowest SHAP values per sample were systematically substituted with candidates selected through BLOSUM62 matrix-guided substitutions and analogous amino acid replacements to generate augmented dataset (Table S4). And dataset was subsequently employed to enhance model discriminative capacity during secondary contrastive learning.

Moreover, a contrastive learning framework was adopted in the second stage to address substrate promiscuity and data scarcity challenges. A key advance of DeepAden is the use of ABP-ESM2 for 27-AA binding pocket site representation and molecule language models for molecule representation. After training, the model successfully captures the intrinsic relationships between pocket sequences and substrate molecules, enabling effective discrimination between substrate and non-substrate compounds (Fig. 6A). Here, the prediction performance of DeepAden was systematically investigated by comparing it with three state-of-the-art A-domains substrate prediction tools [i.e., NRPSpredictor2, PKS/NRPS Analysis and PARAS]. Three independent datasets were used to deliver a fair and rigorous benchmark study. The first dataset, denoted as test, is the test portion of the 786 datasets, which was not used in

training, consisting of 245 A-domains sequences covering common substrate types. Overall, DeepAden achieved the highest values in various multilabel accuracy metrics, including precision (0.925) and recall (0.938), when compared with NRPSpredictor2 and PKS/NRPS Analysis (Fig. 6B). Additionally, DeepAden achieved an F1 score of 0.912, whereas NRPSpredictor2 and PKS/NRPS Analysis had scores of 0.773 and 0.627, respectively. This indicates that DeepAden has higher detection sensitivity and accuracy in the prediction of common substrates. The second independent dataset, denoted as new-90, was a set of experimentally validated results released after DeepAden was trained (MIBiG data v4.0 released in November 2024). DeepAden achieved the highest F1 score (0.754) compared with PARAS and NRPSpredictor2 (Fig. 6C). Notably, PARAS also demonstrated strong performance (F1 score: 0.738). However, it failed to return predictions for sequences with <30% identity to GrsA_Phe(PDB: 1amu), whereas DeepAden consistently provided predictions regardless of sequence divergence (Fig. 2C). The third independent dataset, denoted as few-shot, was a set of detailed structure inference data not included in the 786 datasets, focusing on non-proteinogenic substrate types. Again, DeepAden achieved the highest F1 score (0.822) compared with NRPSpredictor2 and PKS/NRPS Analysis (Fig. 6D). DeepAden exhibits distinct advantages over other tools in predicting non-proteinogenic substrates, which may be largely attributed to its entropy-Shapley-guided data augmentation strategy. For PARAS, we restricted our evaluation to the new-90 dataset, as both the test and few-shot datasets are subsets of its training data. To maintain objectivity and ensure a fair comparison, we specifically employed the new-90 dataset—which none of the models had previously encountered—as the evaluation benchmark. Consequently, we compared the F1 scores of model on three datasets before and after data augmentation. While the performance remained comparable on the test dataset, significant improvements were observed on both the new-90 and Few-shot datasets (Fig. 6E). This improvement can be explained by the effective expansion of sequence data for ABPs with rare substrates through data augmentation, which substantially increased the model's learning frequency and prediction accuracy for ABPs with rare substrates. In contrast, the prediction performance for ABPs with common substrates showed limited improvement due to their ample presence in the 786 datasets. In general, DeepAden improved F1 score respectively by 17.8% (0.7738 to 0.9117), 11.2% (0.6777 to 0.7537), and 79.2% (0.4588 to 0.8222) on the test, new-90 and few-shot datasets, comparing to NRPSpredictor2. These results demonstrate that DeepAden is more precise and reliable than previously developed tools, particularly in predicting non-proteinogenic substrates.

Positioning the orphan BGC of NRPS.

Based on genomic mining and metabolomic analysis of the preserved OsiSh-2 in our laboratory, we identified two classes of nonribosomal peptide compounds from mass spectrometry data: the nyuzenamides and the octaminomycins. while AntiSMASH predicted 11 NRPS-related gene clusters, none of these clusters matched the identified compounds (Fig. S5 and Fig. S6). Using the DeepAden, we identified putative biosynthetic gene clusters (BGCs) for both compound families. To establish the gene

cluster-compound correlation, we employed an A-domain based analysis strategy: first, we analyzed the amino acid building blocks of the target compounds, then systematically compared them with prediction results of DeepAden. It should be noted that, given the potential involvement of various modifying enzymes (e.g., isomerases, methyltransferases) during biosynthesis, this study focuses exclusively on the core scaffold derived from the initially recognized A-domain substrates, without accounting for subsequent modifications that may generate derivative structures. Specifically, nyuzenamide consists of the amino acid sequence Phe-Pro-Leu-Gly-Try-Asn-Thr-Hpg-Val-Gly, and octaminomycin contains Thr-Phe-Leu-Val-Pro-Leu-Tyr(Phe)-Pro(Pip) (Figure 7). For all A-domians in nyuzenamide, substrate prediction of NRPSpredictor2 and PARAS failed to provide complete substrate specificity information, whereas DeepAden accurately predicted the substrate preferences of all A-domains (Figure 7A and 7B). Notably, two A-domains in octaminomycin biosynthesis exhibited distinct substrate promiscuity patterns: one accommodated both proline (Pro) and pipecolic acid (Pip) (M24), while the other accepted either tyrosine (Tyr) or phenylalanine (Phe) (M23). NRPSpredictor2 produced no predictions for two A-domain positions, while PARAS, although accurately predicting all single-substrate cases, failed to detect the multi-substrate features. In contrast, DeepAden not only precisely predicted this feature but also identified both possible substrate options, highlighting its unique advantage in predicting substrate flexibility (Figure 7C and 7D). This provides new insights into the substrate selectivity of nonribosomal peptide synthetases.

Discussion

Nonribosomal peptides (NRPs) offer valuable bioactive compounds and drug leads, with A-domains playing a crucial role in determining the chemical structure of NRPs through their substrate specificity. However, existing prediction methods relying on SCCs (10-AA or 34-AA) derived from the GrsA-Phe A-domain template do not generalize across all enzymes and organisms, resulting in limitations in current bioinformatic tools for predicting the substrate specificity of A-domains. Furthermore, future progress demands the development of robust, plug-and-play models that address both substrate promiscuity and the scarcity of A-domains catalyzing non-proteinogenic substrates, ultimately enabling more accurate prediction of the substrate specificity of A-domains. In this study, we propose DeepAden, a two-stage framework that integrates the precise ABP-GAT model for predicting 27-AA ABPs with the comprehensive capabilities of a cross-modal contrastive learning model, which decodes pocket-substrate interactions by leveraging ABP sequences and substrate molecular fingerprints respectively featurized by ABP-ESM2 and MoLFormer. This framework demonstrates state-of-the-art accuracy and generalizability, particularly excelling in the challenging task of predicting non-proteinogenic substrates. Additionally, by incorporating an XAI strategy, we guide position-specific substitutions for data augmentation during model training and enable systematic investigations into substrate-specific residues across Cys-ABPs and non-Cys-ABPs, thereby advancing the understanding and prediction of the substrate specificity of A-domains in a comprehensive and interpretable manner.

Despite transformative advances in substrate specificity prediction of A-domain enabled by machine learning[9-11, 26], the lack of attempts to interpret these models constrains the mechanistic elucidation of residue functions in ABPs. By constructing an ABP-RF model integrated with TreeSHAP analysis of 27-AA ABP sequences, our study provides global interpretability and identifies critical residues distributed across three functional regions: (1) the substrate entry region (α-helical region, positions 2, 3, 5), where position 2 precisely regulates substrate type through steric effects (Fig 4B). For example, ^{2}Phe is commonly found in amino acid-binding ABPs, whereas ^{2}His is frequently observed in aromatic or carboxylic acid-binding ABPs. ^{3}Asp remains highly conserved in amino acid-binding substrates, recognizing substrate amino groups through salt bridge interactions which has reported previously[27]. In ABP with non-amino acid substrates, position 3 converts to ^{3}Ile, ^{3}Asn, or ^{3}Gly and etc, shifting its function from salt bridge interactions with amino group-containing substrates toward hydrogen bonding networks that accommodate substrates lacking amino moieties[27-28]. Position 5, characterized by high variation in ABP and as a core α-helical residue (Fig. 3B and 4A), which has not been discussed before, potentially modulates the overall helical conformation, thereby influencing the spatial orientation and binding capacity of positions 2 and 3. (2) The ATP-surrounding loop region (positions 15, 18, 20) plays a critical role in substrate binding and binding pocket flexibility. At position 18, aromatic residues such as ^{18}Tyr, ^{18}Phe, or ^{18}Trp form essential π-π stacking interactions with the adenine ring of ATP, stabilizing its binding. Interestingly, in Cys-ABPs, this residue is uniquely replaced by ^{18}Gly, which increases loop flexibility to better accommodate the thiol (-SH) side chains of cysteine substrates. ^{20}Pro maintains a rigid loop conformation and ensures precise alignment with the ribose-phosphate group angles of ATP. In contrast, ^{20}Ala introduces flexibility and allows the loop to adapt to specialized substrates such as Cys. Previous studies primarily focused on the residues surrounding the substrate, while neglecting the overall functional role of the A-domain as an adenylation domain, particularly with respect to the amino acyl binding pocket[29]. Specifically, when comparing Cys-ABP and non-Cys-ABP, earlier research predicted that the 10-AA codes of the A-domain for a cysteine substrate was DLFNLSLIWK, whereas the binding pocket residues actually identified were FDLSGATIWK (PchE numbering)[3]. It has been proposed that two pairs of hydrophobic interaction partners (Ile846/Trp847 and Phe741/Leu743) might confer specificity for cysteine binding by narrowing the binding pocket space[30]. However, a similar Phe/Leu pair was also found in Ala-ABP, and the Ile/Trp pair appeared in Gly-ABP. Our SHAP analysis provides a more comprehensive and reasonable explanation of the mechanism underlying the specificity of A-domain recognition for cysteine. (3) The substrate side chain contact region (positions 21, 24, 25), which directly interacts with substrate side chains, differentiates chemical properties through hydrophobic interactions (^{24}Ile or ^{24}Leu) or hydrogen bonding networks (^{25}Asn or ^{25}Trp). Comparative analysis reveals that positions 2, 3, 5, 7, 12, 13, 17, 20, 24, and 25 serve as critical divergence points for distinguishing between non-amino and amino acid substrates (Fig S7). These positions exhibit significant conservation-divergence patterns across labeled and unlabeled datasets, indicating the presence of abundant, yet undiscovered A-domains with non-amino substrates in metagenomes. This discovery

provides a novel theoretical foundation for natural product mining and the engineered design of proteins.

Although developed A-domain substrate prediction tools have attempted to integrate structural information predicted by AlphaFold to improve accuracy[11, 26], ESM2 has proven to be more efficient in extracting implicit structural features[31]. However, these tools still rely on the traditional approach of using GrsA-Phe (PDB: 1amu) as a template to extract 10-AA and 34-AA codes. Both our analysis and other studies have demonstrated that SCCs extracted based on the binding pocket of A-domains with amino acid substrates fail to address the challenges posed by evolutionary divergence[3]. While AlphaFold can assist in locating SCCs through structural alignment, performing binding pocket analysis on the large-scale datasets involved in our research is infeasible. The 27-AA ABP was created not to replace the 34-AA SCCs but to define a binding pocket within a 6Å range, making it more easily predictable by ABP-GAT using only sequence input and eliminating the heavy reliance on structure predictions. Furthermore, ABP-GAT overcomes the limitations posed by evolutionary divergence, enabling ABP extraction from non-homologous A-domains with 30%-40% sequence similarity to 1amu. This not only enhances prediction specificity for non-proteinogenic substrates, such as aromatic and carboxylic acids, but also facilitates the investigation of ABP residue contributions in our research. Compared to AdenPredictor, DeepAden achieves approximately a 4-fold improvement in average performance with the same amount of training data. While our model performs on par with PARAS, it requires only one-fifth of the training data. This is attributed to our effective data augmentation strategies for biological data and the use of multimodal contrastive learning, which addresses the challenges of training with limited samples. Despite overcoming the impact of sequence evolutionary differences on ABP prediction through PLM-driven structural information, the substrate prediction process fundamentally relies on sequence-based features. For instance, in the new-90 independent test set, performance of DeepAden showed only a 12.9% improvement compared to NRPSpredictor2. To address these challenges, it is necessary to develop multimodal datasets that integrate atomic-level binding pocket features and to design conformation feature extractors based on geometric deep learning algorithms that enable “sequence-to-conformation-to-substrate” prediction[32]. It is worth noting that the 27-AA ABP proposed in this study has already established a solid foundation for achieving this goal. By analyzing the mapping relationships between ABP sequences and conformations, for example, the ^{20}Pro to ^{20}Ala mutation inducing loop flexibility changes, this framework provides valuable guidance for the rational design of substrate binding pockets in synthetic biology. Given the superior performance of DeepAden in predicting non-proteinogenic substrates, we anticipate that this tool will significantly accelerate the discovery of novel NRPSs through genome mining and provide critical technological support for microbial natural product development.

Methods

DeepAden comprises two key components: an A-domains binding pocket prediction model and a substrate specificity prediction model. This section is organized into three main parts: “GAT-based node

classification framework for A-domains binding pocket prediction", "XAI-guided functional residue recognition in A-domain binding pocket" and "Contrastive learning for pocket-substrate binding prediction". Data collection and preprocessing was detailed in Supplementary file.

1. GAT-based node classification model for A-domains binding pocket prediction

DeepAden processes 1D A-domains sequence into 2D graphs for candidate substrate binding pocket predictions. Binding pockets are predefined by calculating the distances between the substrates and the A-domains in 13 co-crystal complexes. Residues are considered part of a binding pocket when the distance is less than 6 Å. These complexes include both proteinogenic and non-proteinogenic substrates (Table S1).

1.1 PLM-based graph construction

For a given 1D amino acids protein sequence with N_p number of residues, DeepAden generates the 2D protein contact map graph with N_p residue nodes using ESM2. The protein graph is characterized by $P^{(0)}$, CM_j^i and $E_p^{(0)}$, representing the residue node features, and the contact map matrix that connects the nodes and the edge features describing these connections, respectively. The residue feature is computed as:

$$P^{(0)} = \mathrm{MLP}_{\mathrm{esm}}\left(P_{\mathrm{esm}}^{(0)}\right) + \mathrm{MLP}_{\mathrm{pc}}\left(P_{\mathrm{pc}}^{(0)}\right) \quad (1)$$

where $P^{(0)} \in \mathbb{R}^{N_p \times 512}$, represents the computed residue feature, $P_{\mathrm{esm}}^{(0)} \in \mathbb{R}^{N_p \times 1280}$ represents the residue-type feature obtained from the final layer of ESM2, and $P_{\mathrm{pc}}^{(0)} \in \mathbb{R}^{N_p \times 18}$ represents physicochemical properties that encodes the residue weights, pK values, hydrophobicity, and whether the residue is aliphatic, aromatic, neutral or charged. The contact map matrix, CM_j^i, is predicted using Residual Neural Network (ResNet) with three residual blocks on the ESM2 attention map, and the edge feature, $E_p^{(0)}$, are obtained from predicted contact probabilities from the ResNet output where each element CM_j^i indicate the probability of contact between residues i and j.

1.2 GAT-based binding pocket prediction model architecture

The GAT model serves as the foundational architecture for node-level classifications on protein graphs, which specifically targets the prediction of binding pocket residues. After constructing the graph, where each node is characterized by an initial feature $P^{(0)}$ and each edge by the feature $E_p^{(0)}$, a graph attention neural network (GAT) is employed to embed the constructed graphs into fixed-size latent representations, typically those residues involved in binding pocket. The core of the model is built upon four GATConv layer. In each layer, the model aggregates information from neighboring nodes using an attention mechanism. Specifically, for the $l-th$ layer, the feature update for node i is defined as:

$$\mathrm{h}_i^{(l+1)} = \mathrm{Dropout}\left(\sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} \mathrm{W}^{(l)} \mathrm{h}_j^{(l)}\right)\right) \quad (2)$$

where $\mathrm{h}_i^{(l)}$ refers to the feature vector of node i at layer l, $\mathrm{W}^{(l)}$ is the learnable weight matrix at layer

l, $\alpha_{ij}^{(l)}$ denotes the attention coefficient between node i and its neighbor j (with $j \in N(i)$), and $\sigma(\cdot)$ is an ELU activation function which is applied after each convolution except in the final layer, and dropout is used to randomly zero out a fraction of the activations to reduce overfitting. The final convolutional layer, refers to as conv4, outputs the logits for each node:

$$z_i = \mathrm{conv4}(h_i^{(3)}) \quad (3)$$

where z_i represents the unnormalized scores for each class. These logits are subsequently transformed into probabilities via the softmax function during loss computation. Due to the limited training data (only 78 crystal PDBs, Supplementary file), we further incorporate data augmentation and consistency regularization. Specifically, Gaussian noise is added to the initial node embeddings to generate perturbed versions of node features:

$$\widetilde{h_i^{(0)}} = h_i^{(0)} + \varepsilon \ (\varepsilon \sim \mathcal{N}(0, \sigma^2 I)) \quad (4)$$

where σ controls the standard deviation of the noise.

To further enhance the robustness of the initial MLP classifier, we employ a consistency regularization based semi-supervised framework, as used by Huang *et al.*, 2023[16b]. The dataset of labeled and unlabeled A-domain structure are described on supplementary file. The overall training loss is defined as:

$$L_{total} = L_{CE} + \lambda L_{cons} \quad (5)$$

where λ is a hyperparameter that balances the contribution of the consistency regularization relative to the cross-entropy loss. Detailed hyperparameter configurations are summarized in Table S5.

2. XAI-guided functional residue recognition in A-domain binding pocket

2.1 A-domain binding pocket residue prediction and sequence analysis

ABP residues, both with and without substrate information, were predicted using ABP-GAT. The predicted ABP residues were concatenated into continuous sequences and pairwise compared to generate a sequence similarity matrix. A sequence similarity network was subsequently constructed using a 70% identity edge cutoff threshold, followed by visualization in Cytoscape v3.9.1. To characterize residue conservation patterns, amino acid frequency matrices were derived from with substrate information ("labeled") and without substrate information ("unlabeled") ABP sequences. Position-specific conservation scores were then quantified using Jensen-Shannon divergence (JSD) to measure deviations in amino acid distributions across alignment positions.

2.2 TreeSHAP model construction for functional and substrate-specific residues analysis

TreeSHAP, a variant of SHAP, leverages Shapley values from cooperative game theory to provide sample-level feature attribution for tree-based models (e.g. random forests, gradient-boosted trees),

offering mathematically consistent and human-interpretable explanations, making it a gold-standard method in XAI applications. To featurize ABP residues, we finetuned ABP-ESM2, which shares the same network architecture and parameter size as the ESM2-650M, using labeled and unlabeled ABP sequences to construct an ABP-specific language model, ABP-ESM2 (Supplementary file). To construct a reliable TreeSHAP model, we first augmented our dataset of 786 ABP sequences (Table S2) to address class imbalance. For each substrate class with fewer than 10 samples, we generated additional samples through random sampling with replacement to ensure a minimum of 10 samples per class, resulting in an augmented dataset of 1579 ABP sequences. Each pocket sequence was tokenized and processed through the ABP-ESM2. For each position in the sequence, we applied UMAP to reduce the dimensionality of the embeddings from 1280D to 1D. The dimensionality-reduced features were used to train a Random Forest classifier using 5-fold cross-validation with GridSearchCV to optimize hyperparameters including n_estimators, max_depth, min_samples_split, min_samples_leaf, and bootstrap options (Fig. S8). For the trained Random Forest model ABP-RF, TreeSHAP calculates the contribution of each feature (sequence position) to the model prediction for each sample and class. The SHAP value for feature i for sample x and class c is defined as:

$$\phi_i(f,x,c) = \sum_{S\subseteq N\setminus\{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!}[f_c(S\cup\{i\}) - f_c(S)] \quad (6)$$

where N is the set of all features, S is a subset of features excluding i, and f_c is the expected prediction for class c when only features in set S are known. We analyzed SHAP values across different classes to identify substrate-specific residues. For each site i, we calculated the mean absolute SHAP value across all classes:

$$\text{SiteImportance}_i = \frac{1}{N\times C}\sum_{j=1}^{N}\sum_{c=1}^{C}\left|\phi_i(f,x_j,c)\right| \quad (7)$$

where C is the number of classes and N is the number of samples. For specific classes of interest, we compared their SHAP profiles with other classes to identify discriminative residues:

$$\text{ClassSpecificImportance}_{i,c} = \frac{1}{|X_c|}\sum_{x_j\in X_c}\phi_i(f,x_j,c) \quad (8)$$

where X_c is the set of samples belonging to class c.

3. Contrastive learning for pocket-substrate binding prediction

3.1 Pocket featurization (pocket encoder)

We generate the substate features of A-domains using ABP-ESM2. To improve the discriminative capacity of pocket encoder for substrate-specific pockets, we adopt the approach of CLEAN by incorporating substrate label information and refining the pocket representation space through supervised contrastive hard loss (SupConHardLoss) in the Euclidean metric space, thereby enhancing pocket feature discriminability[33]. For a protein pocket of length n, the pocket encoder generates a pocket embedding

$P_{full} \in \mathbb{R}^{n \times d_p}$ (d_p=256) which is mean-pooled along the length of the pocket resulting in a vector $P \in \mathbb{R}^{d_p}$. Notably, the pocket encoder is used exclusively to generate pocket features, with their weights keeping unchanged and are not updated during training.

3.2 Substrate featurization (molecule encoder)

We generate the substate features of A-domains using a pretrained CLM MoLFormer, an encoding of the SMILES string of a molecule as a fixed-dimension embedding $M \in \mathbb{R}^{d_m}$(d_m=768). Like the pocket encoder, MoLFormer serves exclusively as a feature generator of molecules, and its parameters are kept frozen throughout the training process.

3.3 Projection and prediction

Given a pocket embedding $P \in \mathbb{R}^{d_p}$and a ligand embedding $M \in \mathbb{R}^{d_m}$, we project them into a shared embedding space ($P^*, M^* \in \mathbb{R}^d$, d=128) using two separate two-layer MLPs:

$$\mathrm{P}^* = \mathrm{MLP}_{\mathrm{pro}}(\mathrm{P}) \quad (9)$$

$$\mathrm{M}^* = \mathrm{MLP}_{\mathrm{mol}}(\mathrm{M}) \quad (10)$$

Where $MLP_{pro}: \mathbb{R}^{d_p} \to \mathbb{R}^d$ and $MLP_{mol}: \mathbb{R}^{d_m} \to \mathbb{R}^d$ are two-layer perceptron with ReLU activation. Given the latent embeddings P^*, M^*in the shared space, we employ cosine similarity as a metric to measure their binding affinity $\hat{p}(P^*, M^*) = cos(P^*, M^*)$. The intuition behind using cosine similarity is that pocket-ligand pairs with stronger binding affinities tend to exhibit higher cosine similarity values and are therefore closer in the embedding space.

3.4 Pocket-substrate pairing pretrained using contrastive learning

The model is trained to predict affinity between pocket and ligand, with the loss computation depending on the training dataset. To enhance cross-modal alignment, we propose a contrastive learning framework that optimizes feature representations by simultaneously attracting genuine binding pairs in the embedding space while repelling non-binding counterparts. We implement a global negative sampling strategy using the entire training dataset. Specifically, for each pocket/molecule pair P_i^*/M_j^*, we consider all other pocket/molecule P_i^*/M_j^* ($i \neq j$) in the complete training set as negative samples, excluding only the known positive binding pair P_i^*/M_j^* ($i = j$). This comprehensive approach ensure that each pocket is contrasted against the full spectrum of molecules in the dataset, providing robust discriminative learning signals across the entire chemical and structural space represented in our training data. Here, we introduce a supervised contrastive loss designed to learn cross-modal alignment between pocket and molecule embeddings through pairwise similarity optimization. Given a batch of pocket-molecule pairs $\{P_i^*, M_j^*\}_{i,j=1}^N$, the similarity matrix is first computed which is derived as:

$$\mathrm{S} = \frac{\hat{\mathrm{p}}(\mathrm{P}^*, \mathrm{M}^*)}{\tau} \quad (11)$$

where τ is a temperature hyperparameter. The loss focuses on the diagonal elements s_{ii} (paired pocket-

molecule similarities) and employ binary cross-entropy (BCE) to jointly discriminate positive and negative pairs:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log \sigma\left(\frac{s_{ii}}{\tau}\right) + (1 - y_i)\log\left(1 - \sigma\left(\frac{s_{ii}}{\tau}\right)\right)\right] \quad (12)$$

where $\sigma(\cdot)$ denotes the sigmoid function, $y_i \in \{0,1\}$ indicates ground-truth pairing labels, and N is the batch size. By leveraging temperature scaling to calibrate confidence in similarity distributions, this loss directly optimizes the discriminative boundaries of paired instances, making it suitable for supervised contrastive learning in tasks such as pocket-molecule interaction prediction. Detailed implementation specifications and parameter configurations are provided in Table S5.

3.5 Substrate assignment and model evaluation

During the inference phase, the model first precomputes binding affinity scores between the query protein pocket and all candidate molecules in the database. The top-10 molecules with the highest predicted binding affinity scores are initially selected as potential binding molecules. Subsequently, we employ a maximum separation method that identifies molecules with the largest binding affinity score gaps relative to other candidates, effectively distinguishing the most promising binding molecules from the candidate pool. To provide statistical confidence measures for these predictions, we incorporate a Gaussian Mixture Model (GMM) into our framework for uncertainty estimation. The GMM consists of two Gaussian components that characterize the empirical distributions of binding affinity scores for positive and negative pocket-ligand pairs. By fitting this two-component GMM to the observed score distributions, we enable probabilistic interpretation of predictions in terms of binding likelihood.

4. Experiment validation

4.1 Actinomycete material, culturing conditions, and DNA preparation.

Streptomyces hygroscopicus OsiSh-2 (GenBank accession no. KU324469.1) was originally isolated from the sheath of healthy rice (*Oryza sativa* cv. Gumei 4) collected in Liuyang, Hunan Province, China. For primary cultivation, the strain was grown on Gauze's agar plates at 28°C for 5 days. Subsequently, a seed culture was prepared by inoculating the strain into liquid Gauze's medium, followed by incubation at 28°C with agitation at 1800 rpm for 5 days. The cultured broth was then transferred to Gauze's plate medium for solid-state fermentation at 28°C for 7 days. High-quality genomic DNA was extracted using a standardized protocol.

4.2 LC-MS/MS analysis and molecular networking

Crude extracts obtained from different fermentation methods were processed and analyzed by LC-DAD-MS. Chromatographic separation was performed using an Agilent G6500 UHPLC system coupled to a quadrupole time-of-flight (Q-ToF) mass spectrometer operated in positive ion mode. Samples (10 μL injection volume) were separated on a ZORBAX Eclipse Plus C18 Rapid Resolution HD column (1.8

μm, 2.1 × 150 mm; Agilent Technologies, USA) maintained at 30 °C with a flow rate of 0.3 mL/min. The mobile phase consisted of (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile (ACN), with the following gradient program: 20% B (0–2 min), linear increase to 100% B (2–14 min), hold at 100% B (14–16 min), followed by re-equilibration at 20% B for 2.8 min (total run time: 18.8 min). The spray chamber conditions were: nebulizer, 5 L/min; drying gas, 200; sheath gas temperature, 350°C, sheath gas flow, 11 L/min; drying gas on, 5 L/min.

Raw LC-MS data from OsiSh-2 fermentation extracts were processed using the Global Natural Product Social Molecular Network (GNPS) platform. Molecular networks were constructed and visualized using Cytoscape (v3.10.1) with the following parameters: precursor and fragment ion mass tolerances of 0.02 Da, minimum cosine score of 0.6, at least two matched fragment ions, and a minimum cluster size of 1. Unidentified metabolites were further annotated by querying the NPAtlas, Reaxys, and SciFinder databases.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32170079 to Z.Q., 32200035 to H.Z., and 32400235 to J.H.), the Natural Science Foundation of Guangdong (2024A1515012593 to Z.Q. and 2023A1515110175 to J.H.), Guangdong Talent Scheme (2021QN020100 to Z.Q.).

Conflict of Interest

The authors declare no competing financial interests.

Author Contributions

Zhiwei Qin and Heqian Zhang designed and supervised the research. Jiaquan Huang, Liangjun Ge and Yaxin Wu performed the bioinformatic and established the algorithm. Jun Wu performed experiment validation. Qiandi Gao and Pan Li participated the bioinformatic analysis. All authors analysed and discussed the data. Zhiwei Qin, Jiaquan Huang and Liangjun Ge wrote the manuscript and all authors edited.

Data Availability Statement

The data and materials supporting the findings reported in this study are publicly available at GitHub repository: https://github.com/Qinlab502/DeepAden.

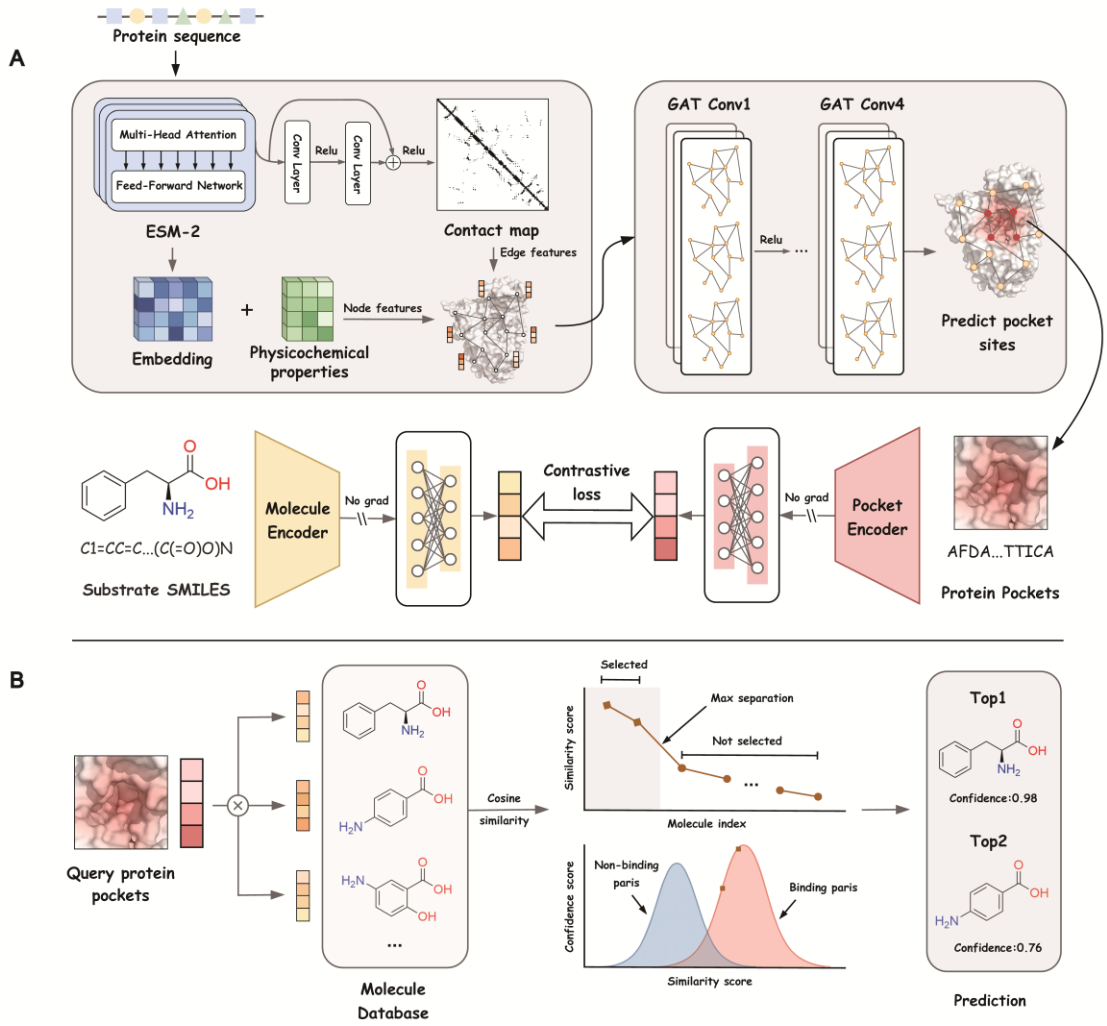

Figure 1. Overview of DeepAden. (A) Training phase of DeepAden. The graph attention neural network-based A-domain binding pocket (ABP) prediction model takes A-domain sequences as input. First, sequence embeddings are extracted using ESM-2, while contact map are derived by integrating attention maps from ESM-2 with a ResNet. These sequence embeddings and physicochemical properties are then used as node features, and residue contact information serves as edge features to construct a graph representation for each sequence. Finally, the graph attention neural network predicts the node positions of the binding pocket. For pocket-substrate binding prediction, a contrastive learning framework is employed. pocket sequence features and substrate SMILES representations are extracted using ABP-ESM2 and MoLFormer, respectively. The contrastive loss between the two modalities is computed to maximize feature similarity. **(B)** Inference phase of DeepAden. The cosine similarity is calculated between the pocket sequence features and the features of all molecules in the substrate database. Candidate molecules are ranked by similarity scores, and the most promising binding molecules are selected using max-separation method. Finally, predicted substrate molecules are assigned confidence scores using a Gaussian Mixture Model (GMM).

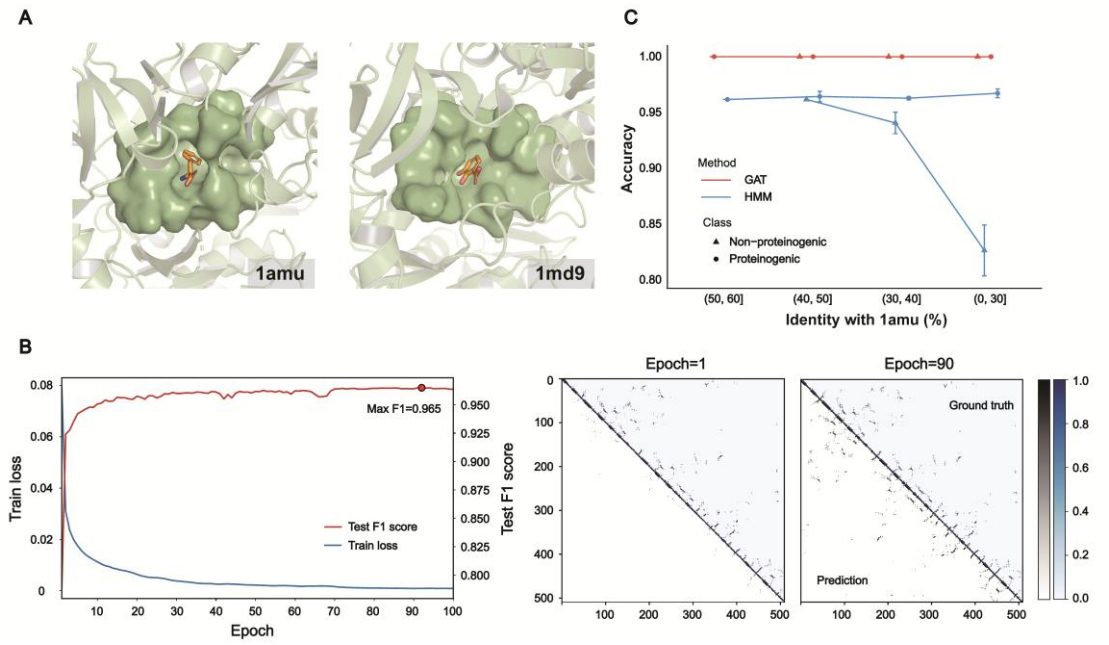

Figure 2. Performance of A-domain binding pocket (ABP) prediction model. (A) The binding pocket was defined by analyzing residues within a 6 Å radius of the substrate, comprising 27 key residue positions. **(B)** Left panel: Training loss of the ResNet model shows continuous reduction during training, while achieving a maximum F1 score of 0.965 on the test set. Right panel: Comparison of contact maps between early (Epoch 1) and late (Epoch 90) training stages, with blue and gray regions representing ground truth and prediction, respectively. **(C)** Accuracy comparison between two methods (GAT and HMM) for predicting binding pocket residues in A-domains recognizing proteinogenic and non-proteinogenic substrates.

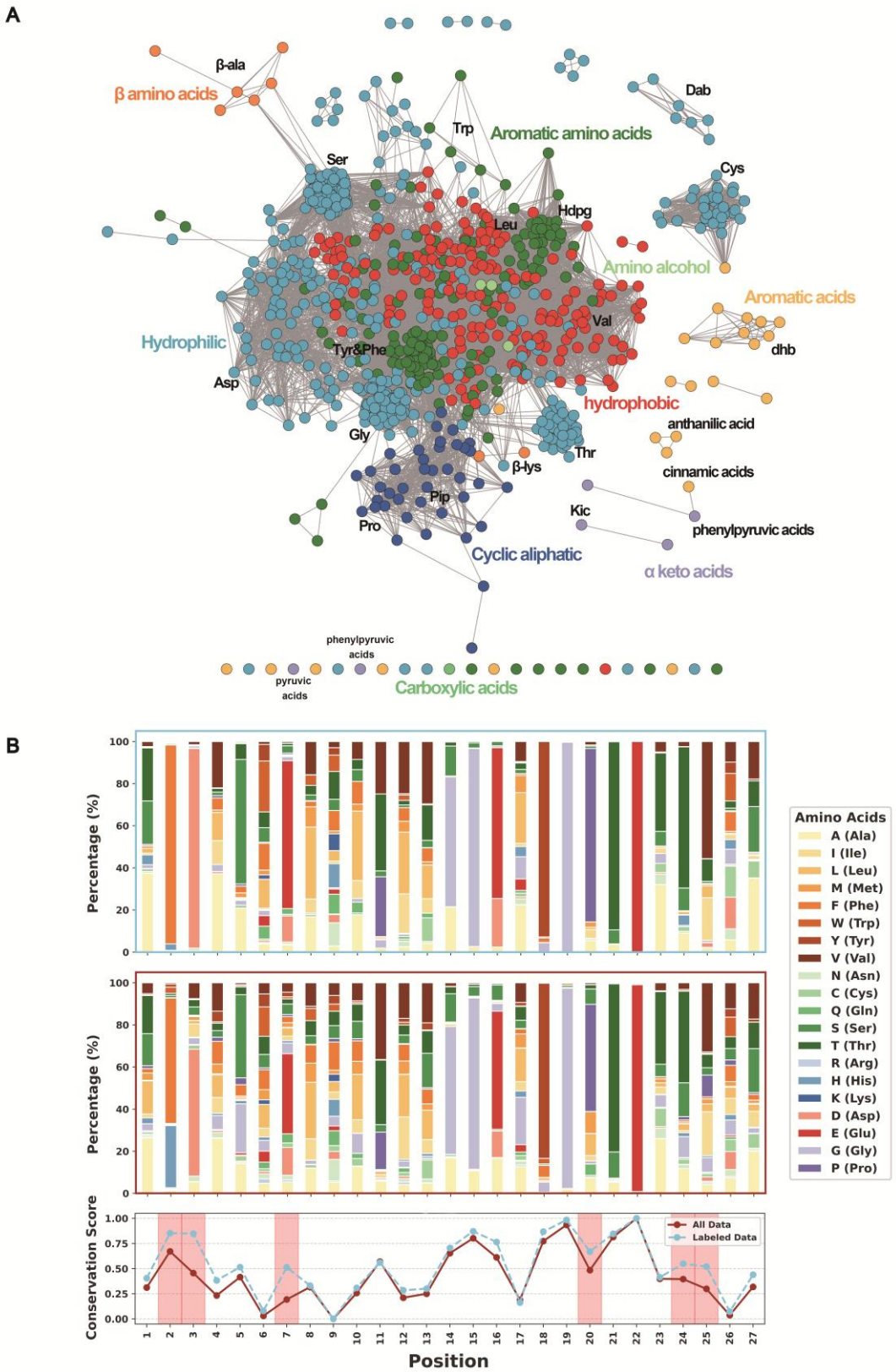

Figure 3. Conservation and similarity analysis of A-domain binding pocket (ABP) sequences. (A) Sequence similarity network of ABP sequences, constructed at a stricter edge detection threshold (~70% identity) and colored according to substrate properties. Subnetworks with the same substrate properties

cluster individually and are labeled accordingly. **(B)** Stacked amino acid frequency plot for labeled ABPs with substrate information and for all ABPs (both labeled and unlabeled). The plot for labeled ABPs is framed with a blue line, while the plot for all data is framed with a red line. Conservation scores at each position were calculated using the Jensen-Shannon divergence method. The red shading in the conservation score line plot highlights differences in conservation scores between the labeled ABPs and all ABPs.

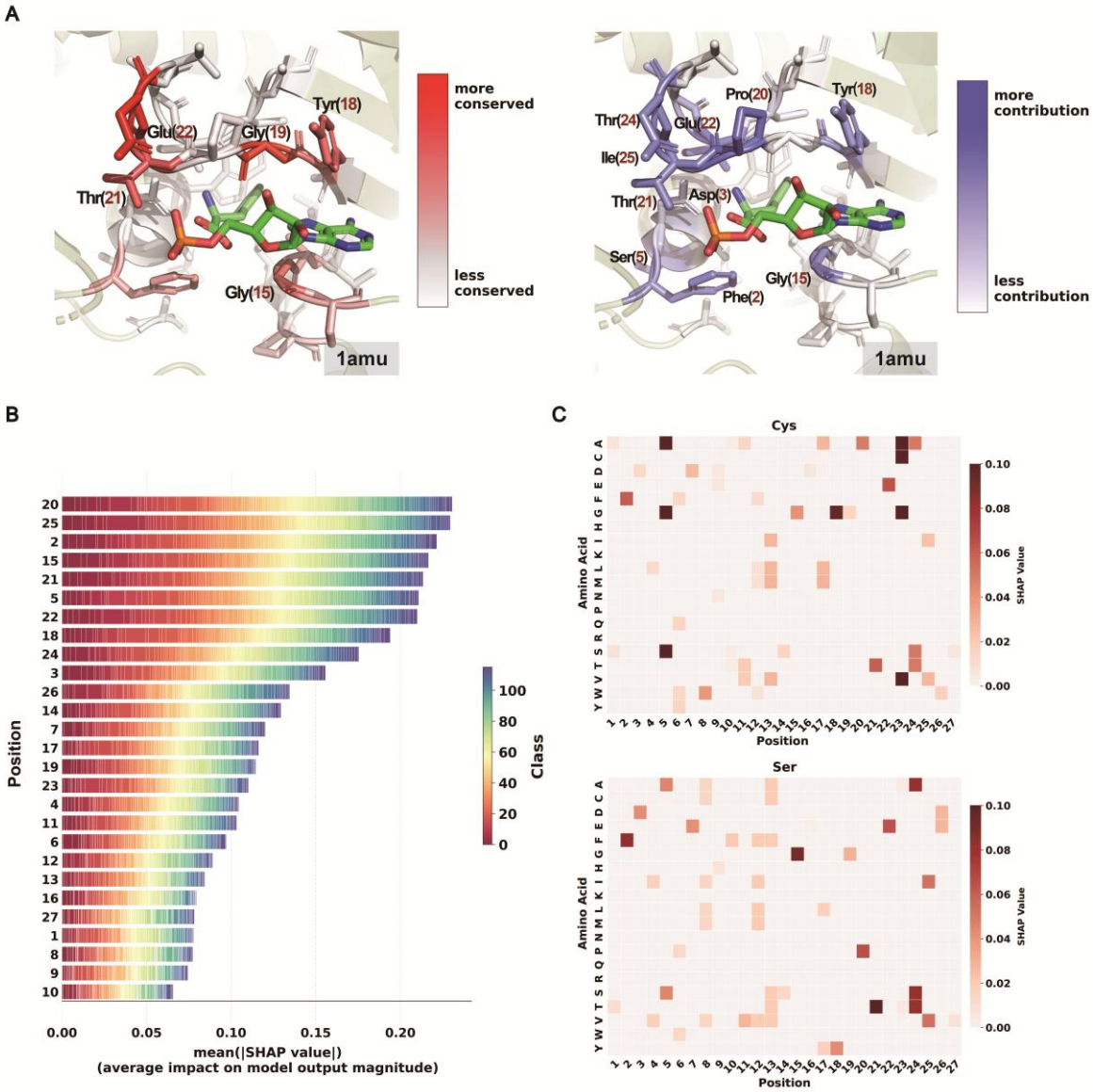

Figure 4. XAI-guided analysis of functional and substrate-specific residues in A-domain binding pocket (ABP) (A) Conservation score (left) and contribution score (right) of each residue in the ABP shown in the crystal structure of GrsA-PheA (PDB:1amu). **(B)** Stacked bar chart of mean |SHAP Value| for 27 amino acid residues in ABPs contributing to the output magnitude of the ABP-RF model. **(C)** Heatmaps of positive SHAP values highlighting amino acid contributions at each residue position for Cys-ABPs and Ser-ABPs within their respective classes.

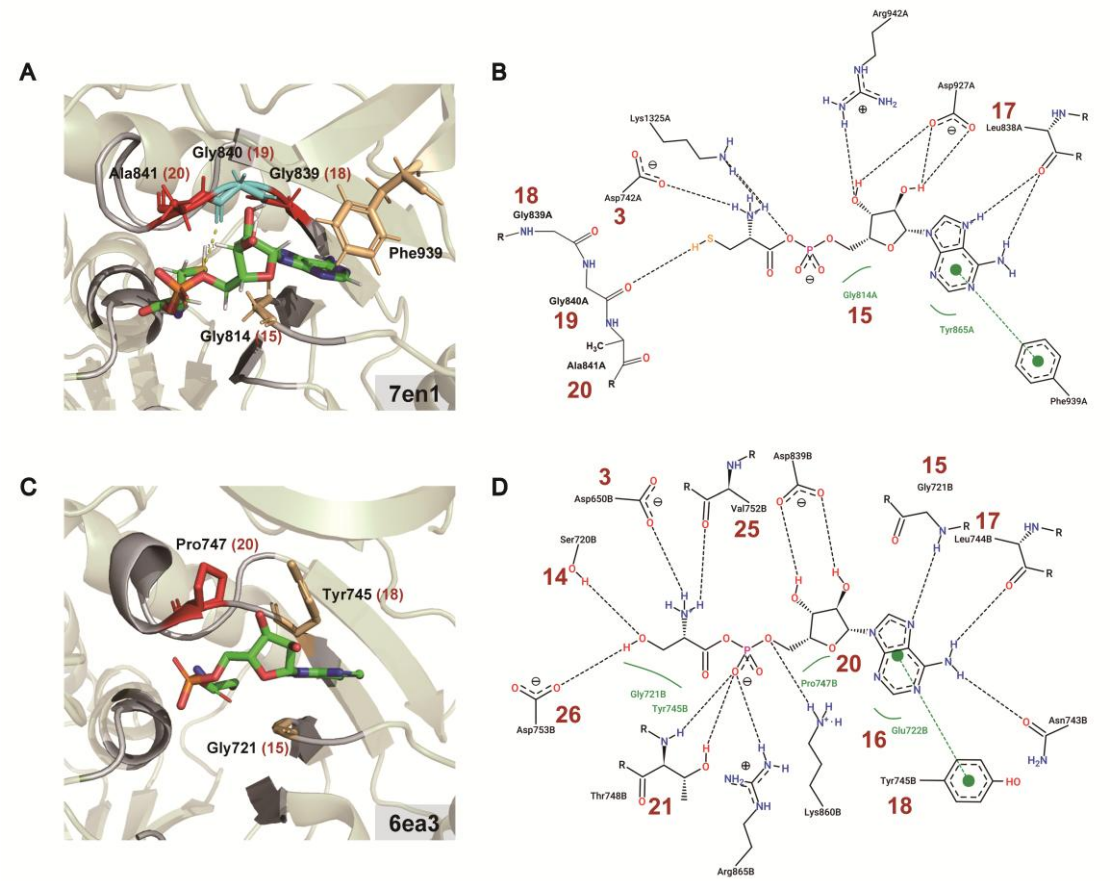

Figure 5. Structural and interaction analysis of functional residues in A-domain binding pocket (ABP) based on PchE-Cys and FscH-Ser crystal structures. (A) Structural representation of functional residues within the ABP as observed in the PchE-Cys crystal structure (PDB: 7en1). The ligand containing adenine ring and pentose sugar is Cys-AMP. The red numbers in parentheses represent positions in the ABP, while black text indicates specific amino acids and their positions in the crystal sequence. **(B)** Two-dimensional diagrams of PchE-Cys (PDB: 7en1) according to chemical drawing conventions. Directed bonds between protein and ligand are drawn as dashed lines, with the interacting protein residues and the ligand visualized as structure diagrams. Hydrophobic contacts are represented more indirectly through spline sections highlighting the hydrophobic parts of the ligand and the label of the contacting residue. Dashed lines between aromatic rings represent π-π stacking interactions. **(C)** Structural representation of functional residues within the ABP as observed in the FscH-Ser crystal structure (PDB: 6ea3). The ligand containing adenine ring and pentose sugar is Ser-AMP. Other annotations follow the same convention as in **(A)**. **(D)** Two-dimensional diagrams of FscH-Ser (PDB: 6ea3) according to chemical drawing conventions. Other annotations follow the same convention as in **(B)**.

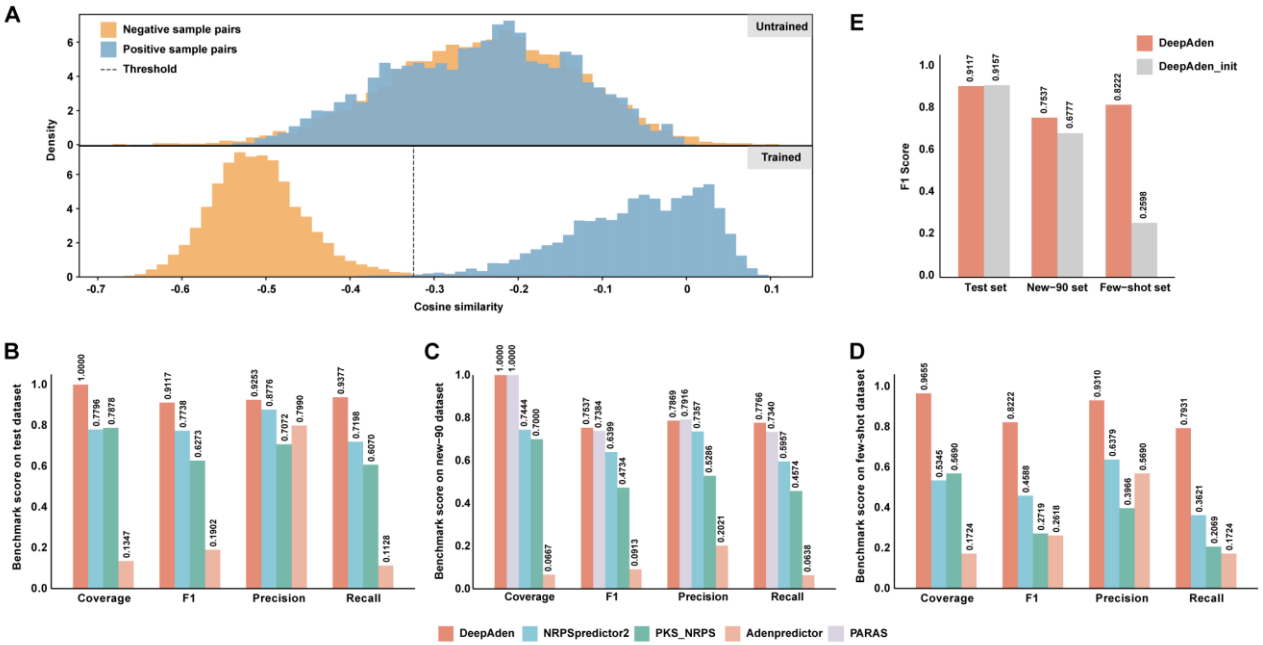

Figure 6. Model performance evaluation of DeepAden. (A) The cosine similarity distribution changes between positive sample pairs (pocket-substrate) and negative sample pairs (pocket-non-substrate) in the training set before and after model training. The separation between positive and negative sample distributions shows significant improvement after training. **(B-D)** Comparative evaluation of DeepAden against state-of-the-art tools (NRPSpredictor2, PARAS, PKS/NRPS analysis, Adenpredictor) across three datasets (test, new-90, and few-shot) using four metrics (coverage, F1-score, precision, and recall). PARAS was compared with DeepAden only on the new-90 dataset, as the test and few-shot datasets overlap with its training data. DeepAden demonstrates superior overall performance compared to other tools across multiple datasets. **(E)** Comparison of F1 score achieved by DeepAden on three datasets before and after implementing data augmentation. Notably, data augmentation leads to a 3-fold improvement in prediction performance for non-proteinogenic substrates.

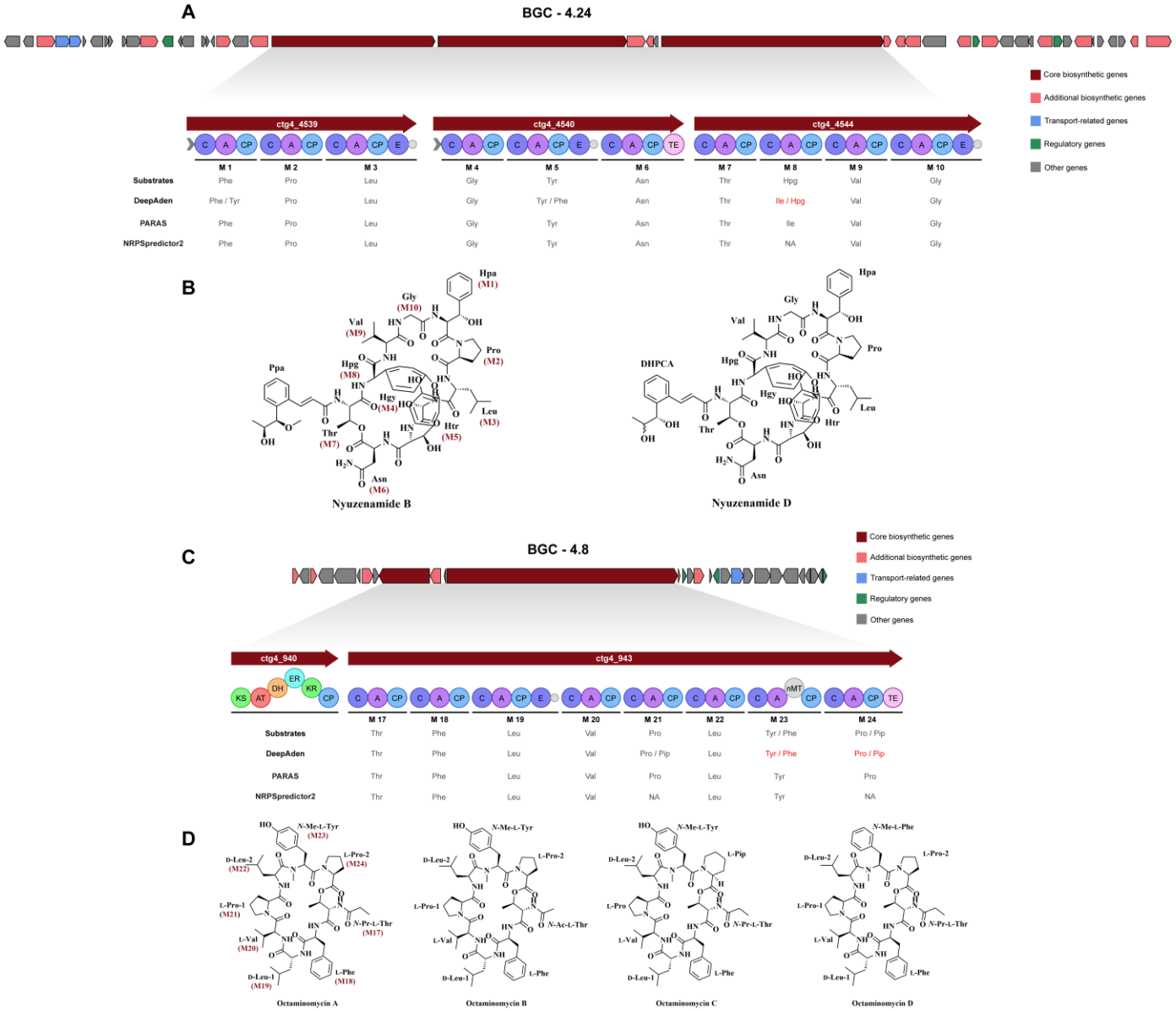

Figure 7. DeepAden demonstrates high prediction accuracy and A-domain substrate promiscuity detection in nyuzenamide and octaminomycin biosynthesis. (A) Precise substrate prediction of DeepAden for all A-domains in the nyuzenamide biosynthetic gene cluster. **(B)** MS-identified nyuzenamide homologs showing their amino acid backbone structures. **(C)** DeepAden profiled all A-domain substrates in the octaminomycin biosynthetic gene cluster with complete accuracy, including correct prediction of promiscuous incorporations (Pro/Pip and Tyr/Phe). **(D)** Four octaminomycin variants resolved by MS, revealing alternative building blocks at two modular positions (M23/M24).

References

[1] R. D. Süssmuth, A. Mainz, *Angewandte Chemie (International ed. in English)* **2017**, *56* (14), 3770, https://doi.org/10.1002/anie.201609079.

[2] R. Finking, M. A. Marahiel, *Annu Rev Microbiol* **2004**, *58*, 453, https://doi.org/10.1146/annurev.micro.58.030603.123615.

[3] S. C. Heard, J. M. Winter, *Natural product reports* **2024**, *41* (7), 1180, https://doi.org/10.1039/d3np00064h.

[4] a) A. Miyanaga, F. Kudo, T. Eguchi, *Current opinion in chemical biology* **2022**, *71*, 102212, https://doi.org/10.1016/j.cbpa.2022.102212; b) A. S. Brown, M. J. Calcott, J. G. Owen, D. F. Ackerley, *Natural product reports* **2018**, *35* (11), 1210, https://doi.org/10.1039/c8np00036k.

[5] T. Stachelhaus, H. D. Mootz, M. A. Marahiel, *Chemistry & biology* **1999**, *6* (8), 493, https://doi.org/10.1016/s1074-5521(99)80082-9.

[6] C. Rausch, T. Weber, O. Kohlbacher, W. Wohlleben, D. H. Huson, *Nucleic Acids Res* **2005**, *33* (18), 5799, https://doi.org/10.1093/nar/gki885.

[7] A. H. Sayed, in (Ed.: A. H. Sayed), Cambridge University Press, Cambridge **2022**.

[8] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, B. Scholkopf, *IEEE Intelligent Systems and their Applications* **1998**, *13* (4), 18, https://doi.org/10.1109/5254.708428.

[9] M. Röttig, M. H. Medema, K. Blin, T. Weber, C. Rausch, O. Kohlbacher, *Nucleic Acids Res* **2011**, *39* (Web Server issue), W362, https://doi.org/10.1093/nar/gkr323.

[10] M. G. Chevrette, F. Aicheler, O. Kohlbacher, C. R. Currie, M. H. Medema, *Bioinformatics* **2017**, *33* (20), 3202, https://doi.org/10.1093/bioinformatics/btx400.

[11] M. Mongia, R. Baral, A. Adduri, D. Yan, Y. Liu, Y. Bian, P. Kim, B. Behsaz, H. Mohimani, *Bioinformatics* **2023**, *39* (39 Suppl 1), i40, https://doi.org/10.1093/bioinformatics/btad235.

[12] F. Kudo, A. Miyanaga, T. Eguchi, *Journal of industrial microbiology & biotechnology* **2019**, *46* (3-4), 515, https://doi.org/10.1007/s10295-018-2084-7.

[13] a) J. P. Nilmeier, E. C. Meng, B. J. Polacco, P. C. Babbitt, in (Ed.: D. J. Rigden), Springer Netherlands, Dordrecht **2017**; b) J. Chen, Z. R. Xie, Y. Wu, *Biochimica et biophysica acta. Proteins and proteomics* **2017**, *1865* (2), 142, https://doi.org/10.1016/j.bbapap.2016.11.008.

[14] a) J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596* (7873), 583, https://doi.org/10.1038/s41586-021-03819-2; b) J. G. Greener, S. M. Kandathil, L. Moffat, D. T. Jones, *Nature reviews. Molecular cell biology* **2022**, *23* (1), 40, https://doi.org/10.1038/s41580-021-00407-0; c) H. Zhang, Y. Wu, Y. Zhu, L. Ge, J. Huang, Z. Qin, *Int J Biol Macromol* **2024**, *265*, 130852, https://doi.org/10.1016/j.ijbiomac.2024.130852.

[15] a) Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. Dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, *Science* **2023**, *379* (6637), 1123, https://doi.org/10.1126/science.ade2574; b) J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, P. Das, *Nat Mach Intell* **2022**, *4* (12), 1256, https://doi.org/10.1038/s42256-022-00580-7.

[16] a) L. Ge, Q. Gao, J. He, X. Wang, J. Huang, H. Zhang, Z. Qin, *Acs Omega* **2025**, *10* (5), 5105, https://doi.org/10.1021/acsomega.4c11266; b) J. Huang, Q. Gao, Y. Tang, Y. Wu, H. Zhang, Z. Qin, *Digit Discov* **2023**, *2* (5), 1484, https://doi.org/10.1039/d3dd00107e.

[17] Q. Gao, L. Ge, Y. Wang, Y. Zhu, Y. Liu, H. Zhang, J. Huang, Z. Qin, *Int J Biol Macromol* **2025**, *285*, 138272, https://doi.org/10.1016/j.ijbiomac.2024.138272.

[18] a) M. R. Karim, T. Islam, M. Shajalal, O. Beyan, C. Lange, M. Cochez, D. Rebholz-Schuhmann, S. Decker, *Brief Bioinform* **2023**, *24* (5), https://doi.org/10.1093/bib/bbad236; b) R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Y. Wen, T. Shah, G. Morgan, R. Ranjan, *Acm Computing Surveys* **2023**, *55* (9), https://doi.org/10.1145/3561048.

[19] a) Q. Dickinson, J. G. Meyer, *PLoS Comput Biol* **2022**, *18* (1), e1009736, https://doi.org/10.1371/journal.pcbi.1009736; b) S. M. Lundberg, S. I. Lee, presented at *Adv Neur In*, Long Beach, California, USA, **2017**.

[20] H. Kaur, H. S. Pannu, A. K. Malhi, *Acm Comput Surv* **2019**, *52* (4), https://doi.org/10.1145/3343440.

[21] C. T. Walsh, R. V. O'Brien, C. Khosla, *Angewandte Chemie (International ed. in English)* **2013**, *52* (28), 7098, https://doi.org/10.1002/anie.201208344.

[22] a) R. Singh, S. Sledzieski, B. Bryson, L. Cowen, B. Berger, *Proc Natl Acad Sci U S A* **2023**, *120* (24), e2220778120, https://doi.org/10.1073/pnas.2220778120; b) B. Gao, B. Qiang, H. Tan, M. Ren, Y. Jia, M. Lu, J. Liu, W. Ma, Y. Lan, presented at *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, LA, USA, **2023**.

[23] Z. Yang, J. Wang, Y. Zhu, in *Computer Vision – ECCV 2022* (Eds.: S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, T. Hassner), Springer Nature Switzerland, **2022**, 293-309.

[24] T. Schwecke, K. Göttling, P. Durek, I. Dueñas, N. F. Käufer, S. Zock-Emmenthal, E. Staub, T. Neuhof, R. Dieckmann, H. von Döhren, *Chembiochem* **2006**, *7* (4), 612, https://doi.org/10.1002/cbic.200500301.

[25] T. V. Lee, L. J. Johnson, R. D. Johnson, A. Koulman, G. A. Lane, J. S. Lott, V. L. Arcus, *J Biol Chem* **2010**, *285* (4), 2415, https://doi.org/10.1074/jbc.M109.071324.

[26] a) B. R. Terlouw, C. Huang, D. Meijer, J. D. D. Cediel-Becerra, M. L. Rothe, M. Jenner, S. Zhou, Y. Zhang, C. D. Fage, Y. Tsunematsu, G. P. van Wezel, S. L. Robinson, F. Alberti, L. M. Alkhalaf, M. G. Chevrette, G. L. Challis, M. H. Medema, *bioRxiv* **2025**, 2025.01.08.631717, https://doi.org/10.1101/2025.01.08.631717; b) A. K. Adduri, A. T. McNutt, C. N. Ellington, K. Suraparaju, N. Fang, D. Yan, B. Krummenacher, S. Li, C. Bodden, E. P. Xing, B. Behsaz, D. Koes, H. Mohimani, *bioRxiv* **2025**, https://doi.org/10.1101/2025.01.13.632878.

[27] E. Conti, T. Stachelhaus, M. A. Marahiel, P. Brick, *Embo J* **1997**, *16* (14), 4174, https://doi.org/10.1093/emboj/16.14.4174.

[28] a) D. G. Fujimori, S. Hrvatin, C. S. Neumann, M. Strieker, M. A. Marahiel, C. T. Walsh, *Proc Natl Acad Sci U S A* **2007**, *104* (42), 16498, https://doi.org/10.1073/pnas.0708242104; b) J. J. May, N. Kessler, M. A. Marahiel, M. T. Stubbs, *Proc Natl Acad Sci U S A* **2002**, *99* (19), 12120, https://doi.org/10.1073/pnas.182156699.

[29] A. M. Gulick, *ACS Chem Biol* **2009**, *4* (10), 811, https://doi.org/10.1021/cb900156h.

[30] J. Wang, D. Li, L. Chen, W. Cao, L. Kong, W. Zhang, T. Croll, Z. Deng, J. Liang, Z. Wang, *Nat Commun* **2022**, *13* (1), 592, https://doi.org/10.1038/s41467-022-28284-x.

[31] a) Z. Zhang, H. K. Wayment-Steele, G. Brixi, H. Wang, D. Kern, S. Ovchinnikov, *Proc Natl Acad Sci U S A* **2024**, *121* (45), e2406285121, https://doi.org/10.1073/pnas.2406285121; b) H. Y. Koh, A. T. N. Nguyen, S. Pan, L. T. May, G. I. Webb, *Nature Machine Intelligence* **2024**, *6* (6), 673, https://doi.org/10.1038/s42256-024-00847-1.

[32] a) W. Feng, L. Wang, Z. Y. Lin, Y. H. Zhu, H. Wang, J. Q. Dong, R. Bai, H. T. Wang, J. L. Zhou, W. Peng, B. Huang, W. B. Zhou, *Nat Mach Intell* **2024**, *6* (1), 62, https://doi.org/10.1038/s42256-023-00775-

6; b) J. Lu, X. Chen, S. Z. Lu, C. Shi, H. Guo, Y. Bengio, J. Tang, *arXiv* **2025**, *2410.18403*, https://doi.org/10.48550/arXiv.2410.18403.
[33] T. Yu, H. Cui, J. C. Li, Y. Luo, G. Jiang, H. Zhao, *Science* **2023**, *379* (6639), 1358, https://doi.org/10.1126/science.adf2465.