PeptiVerse: A Unified Platform for Therapeutic Peptide Property Prediction

Yinuo Zhang[1,2], **Sophia Tang**[1], **Tong Chen**[1],
Elizabeth Mahood[3], **Sophia Vincoff**[3], **Pranam Chatterjee**[1,3,†]
[1]Department of Computer and Information Science, University of Pennsylvania
[2]Duke-NUS Medical School
[3]Department of Bioengineering, University of Pennsylvania

[†]Corresponding author: pranam@seas.upenn.edu

Abstract

Therapeutic peptides combine the advantages of small molecules and antibodies, offering target flexibility and low immunogenicity, yet their successful translation requires careful evaluation of multiple developability properties beyond binding alone. As chemically modified peptides become increasingly common in drug design, no unified platform currently supports systematic property assessment across both canonical sequences and SMILES-based representations. Leveraging the generalizability of large foundational models trained on protein and chemical data, we introduce **PeptiVerse**, a universal therapeutic peptide property prediction platform. PeptiVerse accepts either amino acid sequences or chemically modified peptide SMILES, delivers state-of-the-art performance across diverse property prediction tasks, and provides both a web interface and open-source implementation for rapid, accessible, and scalable peptide developability analysis. By unifying property prediction across representations, PeptiVerse directly supports early-stage peptide therapeutic development campaigns and property-aware generative design workflows.

1 Introduction

Peptide-based therapeutics have gained significant attention in recent years, highlighted by the clinical success of GLP-1 receptor agonists for metabolic diseases (Zheng et al., 2024; Drucker, 2025). As a therapeutic modality, peptides occupy a unique position between small molecules and antibodies, combining larger interaction surfaces capable of engaging protein-protein interfaces traditionally considered undruggable (Chen et al., 2025a; Wang et al., 2022) with reduced immunogenicity and manufacturing complexity relative to full-length antibodies (Wang et al., 2022; Tang et al., 2025a; Chen et al., 2023). These features make peptides candidates for a broad spectrum of therapeutic targets, including receptors, enzymes, and intrinsically disordered regions.

Despite these advantages, native peptides often exhibit suboptimal translational profiles. For example, poor membrane permeability limits cellular uptake and oral bioavailability (Storchmannová et al., 2025; Chen et al., 2022), rapid proteolytic degradation results in short circulating half-life and frequent dosing (Werle and Bernkop-Schnürch, 2006; Wang et al., 2022), and low solubility promotes aggregation and formulation failure (Malavolta et al., 2006; Zapadka et al., 2017). In addition, certain features of amphipathic sequences can cause hemolysis through nonspecific membrane disruption (Kellermeyer, 1962; Timmons and Hewage, 2020), while nonspecific protein adsorption (fouling) reduces effective concentration and increases off-target interactions in complex biological environments (Jiang and Cao, 2010; Guntuboina et al., 2023). These limitations can be partially mitigated through cyclization, terminal modifications, D-amino acids, or other noncanonical residues (Wang et al., 2022; Li et al., 2023; DeGruyter et al., 2017), but such interventions push peptides beyond the assumptions of traditional protein sequence-based predictors (Notin et al., 2023; Xie et al.,

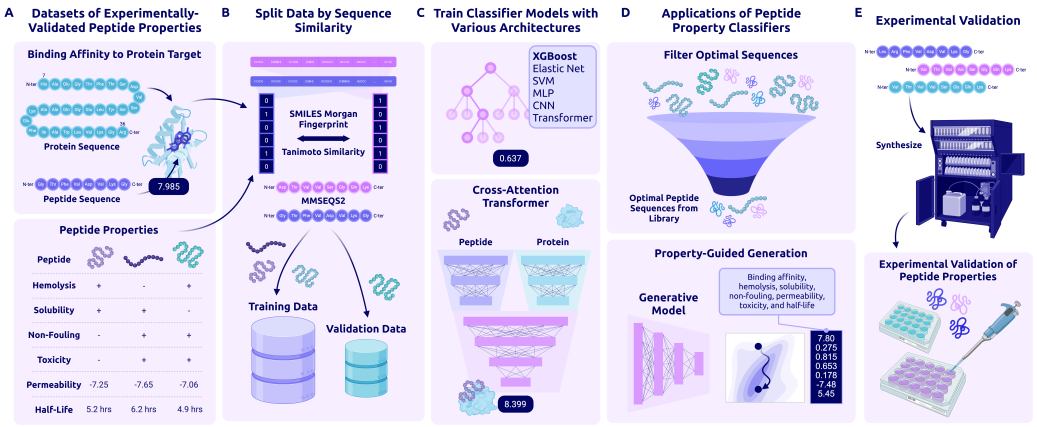

Figure 1: **Overview of PeptiVerse Workflow and Applications. (A)** Dataset collection **(B)** Data splitting **(C)** Property prediction **(D)** Applications **(E)** Downstream validation

2025). As a result, practical peptide design requires systematic evaluation of multiple experimentally grounded properties beyond binding affinity alone.

Existing computational tools inadequately address this need. Sequence-based predictors such as PeptideBERT and those in Peptipedia v2.0 are confined to natural amino acids and cannot accommodate chemical modifications (Guntuboina et al., 2023; Cabas-Mora et al., 2024). Small-molecule ADMET platforms accept "chemical" inputs in the form of SMILES but are trained on drug-like chemical space that differs substantially from peptides and proteins (Fu et al., 2024; Swanson et al., 2024). Peptide-oriented SMILES predictors, including PepLand, PepDoRA, and PeptideDashboard, represent important steps toward chemistry-aware peptide modeling (Zhang et al., 2025a; Wang et al., 2024; Ansari and White, 2023), yet cover only a limited subset of relevant properties. Fitness-focused platforms such as ProteinGym benchmark mutational effects but do not target critical biochemical traits relevant for therapeutic peptide design (Notin et al., 2023; Xie et al., 2025). Together, these gaps highlight the lack of a comprehensive, modality-flexible predictor capable of evaluating the full chemical and sequence diversity of modern peptide therapeutics.

The recent emergence of large protein and chemical language models enables unified property predictors that leverage rich learned representations without requiring explicit structural information (Lin et al., 2023; Feller and Wilke, 2025a). Such predictors have become integral to modern generative peptide design workflows, where classifiers guide sampling, rank candidates, or provide post hoc filtering (**?**Chen et al., 2025b;c; Goel et al., 2025; Tang et al., 2025b;c;d; Zhang et al., 2025b; Vincoff et al., 2025). In this setting, gradient coupling between generator and predictor is not always necessary, allowing effective use of classical methods such as SVM (Hearst et al., 1998), Elastic Nets (Simoncini, 2005), and XGBoost (Chen et al., 2015), which perform well on pretrained embeddings while reducing overfitting risk (De Landsheere et al., 2025).

Here, we introduce **PeptiVerse** (Figure 1), a universal therapeutic peptide property evaluation framework designed to standardize and accelerate computational peptide design. PeptiVerse integrates state-of-the-art foundational models with carefully curated datasets to deliver fast, accurate, and scalable property predictions, supporting both sequence-based and SMILES-based peptide inputs. Beyond post hoc evaluation, these predictors can be directly used as guidance oracles within generative modeling workflows (Tang et al., 2025b;c;d; Chen et al., 2025b;c), enabling generation, ranking, and optimization of peptide candidates across diverse targets. Together, PeptiVerse provides a unified foundation for property-aware peptide discovery, enabling both early-stage candidate prioritization and integration with generative design workflows to accelerate therapeutic translation.

2 Methods

2.1 Data Collection and Preparation

Throughout this work, we distinguish peptide inputs by representation modality rather than biological origin. "Amino acid" inputs refer to canonical sequence-based representations processed by protein language models, while "SMILES" inputs refer to chemistry-aware molecular representations, which may include both canonical and non-canonical peptides. Detailed per-category data collection and model training are discussed in appendix S1.2 S1.1.

3 Results

3.1 Dataset composition highlights constraints across peptide properties

Accurate peptide property prediction is fundamentally constrained by data availability, coverage, and experimental diversity. As an initial step, we curated and integrated experimentally derived datasets spanning multiple peptide properties from a wide range of public sources (Guntuboina et al., 2023; Zhang et al., 2025a; Li et al., 2023; Pirtskhalava et al., 2021; Rathore et al., 2024; Smialowski et al., 2012; Jain et al., 2024; Mathur et al., 2016; D'Aloisio et al., 2021). These datasets collectively cover both canonical amino acid sequences and chemically modified peptides represented as SMILES, enabling unified analysis across representation modalities. Examination of dataset composition reveals substantial heterogeneity in dataset size, label balance, and value distributions across properties (Figure 2; Table S1).

Our data show that several classification tasks are supported by large, well-populated datasets with broad coverage of peptide sequence and chemical space. Hemolysis and non-fouling datasets comprise thousands to tens of thousands of peptides curated from antimicrobial and surface-interaction studies (Pirtskhalava et al., 2021; Barrett et al., 2018), while solubility datasets aggregate protein expression outcomes from structural genomics pipelines and mutational databases (Smialowski et al., 2012; Veleckỳ et al., 2022). Permeability datasets include both canonical cell-penetrating peptides and noncanonical cyclic peptides measured using PAMPA and Caco-2 assays (Avdeef, 2005; Artursson et al., 2012; Li et al., 2023), yielding relatively balanced class distributions and continuous-valued measurements spanning multiple orders of magnitude. These properties exhibit broad value distributions and sufficient sample sizes to support robust model training and evaluation under similarity-aware splits.

In contrast, regression tasks such as peptide half-life and binding affinity remain comparatively data-limited. Half-life measurements, curated from THPdb2, PEPlife, and PepTherDia (Jain et al., 2024; Mathur et al., 2016; D'Aloisio et al., 2021), are sparse, heterogeneous in experimental protocol, and often reported in coarse or qualitative units, resulting in limited sample sizes for both sequence- and SMILES-based representations. Binding affinity datasets aggregate diverse experimental readouts (K_d, K_i, and IC_{50}) across protein–peptide pairs (Zhang et al., 2025a; Lei et al., 2021), but remain modest in scale relative to classification tasks. Together, our observations highlight that achievable predictive performance across peptide properties is frequently constrained by data availability and experimental variability rather than model capacity alone, motivating property-specific modeling strategies and emphasizing the importance of continued dataset expansion.

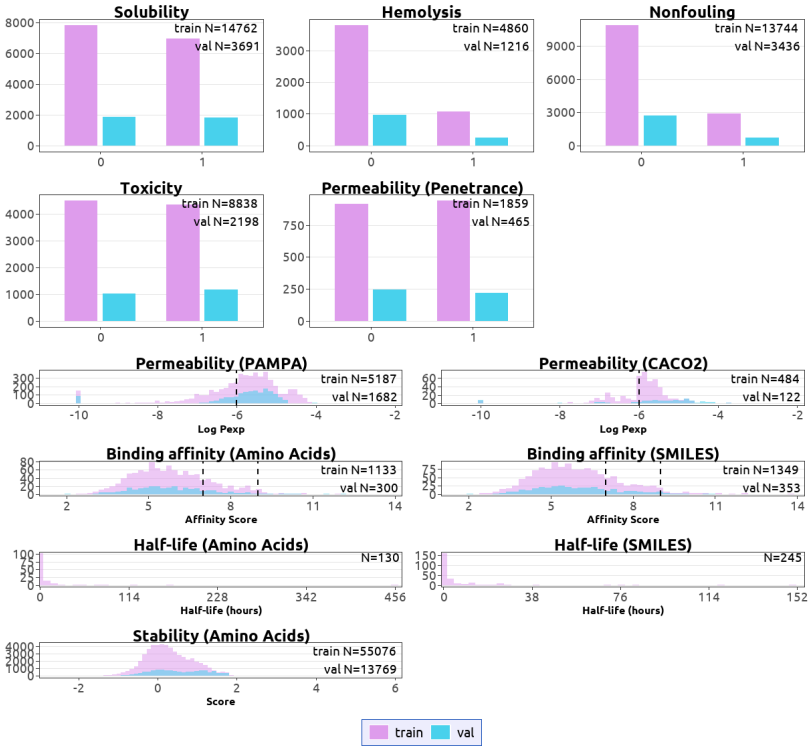

Figure 2: **Training data distributions.** Top: Distributions of binary-labeled datasets across training and validation splits. Bottom: Distributions of continuous-valued datasets across training and validation splits. Purple denotes training samples and blue denotes validation samples. Dashed vertical lines indicate heuristic thresholds for favorable properties, including good permeability ($x = -6$) and strong binding affinity ($x = 7$ and 9). Due to limited data availability, half-life predictors were trained using cross-validation and therefore are shown without an explicit train/validation split.

3.2 PeptiVerse deploys state-of-the-art predictors for a broad range of therapeutically relevant peptide properties

Given our heterogeneous data settings, we evaluated a diverse set of predictor architectures within PeptiVerse, including linear models, boosting methods, multilayer perceptrons (MLPs), convolutional neural networks (CNNs), support vector machines (SVMs), and transformer-based models (Figure 3). Across all classification tasks and both input modalities for amino acid sequence inputs and SMILES inputs, overall performance differences between architectures were modest when trained on fixed, information-rich embeddings derived from ESM-2, PeptideCLM or ChemBERTA (Supplementary Figure S1; Supplementary Tables S2, S3, S4). This consistency suggests that representation quality and dataset characteristics, rather than downstream model capacity, constitute the primary performance bottleneck in peptide property prediction.

Rather than enforcing a single architecture, PeptiVerse identifies the best-performing model per property and representation, reflecting the observation that no single architecture emerges as universally optimal. Different peptide properties favored different inductive biases, with hemolysis and permeability benefiting from margin-based or kernel methods, solubility favoring convolutional architectures, and chemically rich SMILES-based tasks often selecting transformer-based heads (Figure 3). Importantly, similar trends were observed for regression tasks, where multiple nonlinear models achieved comparable performance, and no clear architectural dominance was observed (Supplementary Figure S2; Supplementary Figure S3). In both settings, properties supported by larger datasets with broader coverage of the underlying value distributions, such as the permeability_CPP (cell penetrance ability) task, consistently yielded stronger performance than properties with limited or sparsely

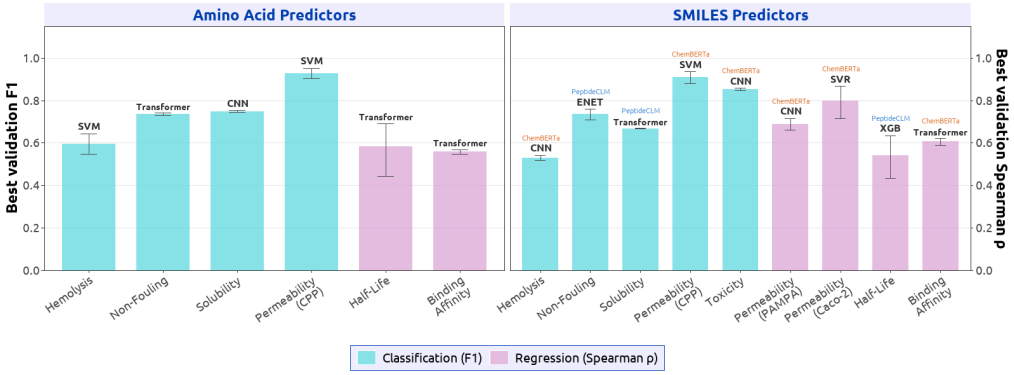

Figure 3: **Best-model performance of PeptiVerse across different therapeutic peptide properties.** Left: Performance on amino acid encoded sequences using ESM-2 embeddings; Right: Performance on SMILES represented sequences using PeptideCLM and ChemBERTa embeddings. Bars report validation set performance of the best-performing model selected after OPTUNA optimization, with names of the model annotated on top. Cyan indicates the performance of the classification models, and light purple indicates the performance of the regression models.

distributed data, such as peptide half-life. Importantly, all evaluations were conducted using similarity-based data splits, indicating that simple models can generalize comparably to deep architectures on out-of-distribution peptide designs when initialized with high-quality representations.

Across all tasks, embedding choice was the dominant factor governing predictive accuracy rather than model architecture. The spread in performance across embeddings consistently exceeded the spread across models within the same embedding. For regression, ChemBERTa consistently outperformed PeptideCLM on both PAMPA ($\rho = 0.69$ vs. 0.59) and Caco-2 ($\rho = 0.80$ vs. 0.75; Supplementary Table S3), likely reflecting ChemBERTa's broader pretraining chemical diversity compared to PeptideCLM's focus on synthetic cyclic peptides. DNN models have low variance with different random seeds, confirming that performance differences across configurations reflect genuine embedding and task effects rather than optimization noise. Although we find that baseline embeddings (e.g. ECFP, VHSE, or one-hot) can achieve competitive performance at times, embeddings from language models consistently attain the highest performance across properties (Supplementary Tables S5, S6, S7, S8, S9, S10). We find that ESM-C embeddings only improve upon ESM-2 embeddings for solubility (F1 = 0.765 vs. 0.754) and halflife (Spearman's rho = 0.636 vs. 0.582; Supplementary Table S11), highlighting ESM-2's competitive representation ability.

3.3 PeptiVerse provides accurate binding affinity predictions

To assess whether PeptiVerse binding affinity predictions reflect physically meaningful interaction strength, we examined their relationship to structure-based confidence metrics derived from state-of-the-art *de novo* protein-peptide complex predictors. Recent studies have reported correlations between structure prediction confidence scores, such as ipTM, and docking-based interactions metrics for protein-protein interactions, motivating the use of such scores as proxies for binding strength in *ab initio* modeling pipelines (Zhu et al., 2023; Peng et al., 2025). We therefore asked whether similar relationships hold for peptide-protein interactions, including both wild-type and chemically modified peptides.

Using OpenFold3-predicted complexes (The OpenFold3 Team, 2025), we compared ipTM scores against experimentally measured binding affinities for 1,433 amino acid sequence inputs and 1,702 SMILES-based peptide inputs. In contrast to prior observations demonstrating the efficacy of the ipTM score for estimating protein-protein interaction strength (Peng et al., 2025), ipTM showed negligible association with experimental binding affinity across either peptide representation (Supplementary Figure S4; $|\rho| \approx 0.1$), indicating that structure confidence metrics are insufficient proxies for peptide–protein binding strength. These results suggest that ipTM does not reliably capture the energetic or kinetic determinants governing

Table 1: **Results for property prediction benchmarks.**

The results reported for PepLand are taken directly from the original publication (Zhang et al., 2025a), which uses random data splits, whereas PeptiVerse employs similarity-aware splits to evaluate out-of-distribution generalization. AA denotes amino acid sequence inputs. Results for PeptiVerse were taken from the Table S2-S3,S6 with uncertainty analysis reported.

	Regression (Spearman Correlation)			Classification (Best F1)	
Method	Binding Affinity (AA)	Permeability (all SMILES combined)	Binding Affinity (SMILES)	Permeability_CPP	Solubility
ESM-2	0.452	–	–	0.885	0.725
PepLand	0.503	0.628	**0.768**	0.838	0.662
PepLand + ESM-2	0.454	–	–	0.885	0.730
PeptiVerse	**0.557**	**0.744**	0.586	**0.929**	**0.749**

peptide binding, likely due to the increased flexibility, shallow binding interfaces, and diverse chemistries characteristic of peptide ligands. Moreover, even when accurate complex structures are available, structure prediction is computationally heavier than embedding-based inference and is not well-suited for high-throughput screening across large peptide libraries. PeptiVerse therefore provides a fast affinity surrogate that complements structure prediction rather than replacing it.

In contrast, PeptiVerse binding affinity predictions show statistically significant agreement with experimental measurements across both representation modalities, achieving Spearman $\rho = 0.55$ for amino acid sequence inputs and $\rho = 0.61$ for SMILES-based inputs (all $p < 10^{-3}$; Supplementary Figure S3). These results, obtained using a transformer-based cross-attention architecture operating directly on protein and peptide embeddings, demonstrate that PeptiVerse provides experimentally relevant binding affinity estimates that complement structural modeling and enable binding-aware filtering and generative peptide design.

3.4 PeptiVerse demonstrates superior predictive performance relative to existing peptide property predictors

PeptiVerse is, to our knowledge, the first unified framework capable of predicting multiple physicochemical and developability properties for both amino acid sequence inputs and SMILES-encoded peptide inputs. In contrast, prior tools have made meaningful progress but remain specialized in either input modality or property scope (Feller and Wilke, 2025a; Zhang et al., 2025a). PeptideBERT (Guntuboina et al., 2023) operates primarily on canonical amino acid sequences and does not generalize to chemically modified peptides. None of the existing tools provides the breadth of properties, multimodality, or user accessibility offered by PeptiVerse.

To quantitatively assess these differences, we compare PeptiVerse against representative prior methods across multiple regression and classification benchmarks in Table 1. Across most evaluated tasks, PeptiVerse achieves competitive or superior performance under a unified evaluation protocol. We note that PepLand reports higher performance on SMILES-based binding affinity prediction (Zhang et al., 2025a). However, these results were obtained using random data splits, which permit substantial overlap in sequence or chemical similarity between the training and test sets. In contrast, PeptiVerse employs similarity-based splits, resulting in a substantially more challenging and realistic evaluation of out-of-distribution generalization. Performance differences on SMILES-based binding affinity, therefore, reflect differences in the evaluation protocol rather than representational limitations.

Additionally, the PepLand+ESM-2 setting is reported only for canonical peptide tasks, and the mechanism for switching between SMILES and sequence inputs is not explicitly specified. By contrast, PeptiVerse provides an explicit and unified multimodal design, enabling consistent evaluation across both canonical and non-canonical peptide spaces.

3.5 PeptiVerse is deployed as a unified web interface for peptide evaluation

Finally, we developed an interactive web interface to make PeptiVerse accessible for practical use by both experimental and computational researchers. PeptiVerse is deployed as a web server (Figure 4) built with `Gradio` (Abid et al., 2019) and hosted on HuggingFace

Spaces (Jain, 2022). The interface allows users to submit peptide inputs as either amino acid sequences or SMILES strings, enabling property evaluation across both canonical and chemically modified peptides. It supports a broad set of therapeutically relevant properties, including binding affinity to a target protein sequence, hemolytic activity, non-fouling behavior, permeability, solubility, toxicity, and peptide half-life.

To improve interpretability and usability, the interface also provides visualizations of the underlying training data distributions, allowing users to contextualize input peptides relative to experimentally characterized datasets. All datasets used to train the deployed models are standardized and distributed in HuggingFace Dataset format, facilitating reproducibility, benchmarking, and integration into downstream peptide design and optimization workflows.

4 Discussion

PeptiVerse introduces a unified framework for therapeutic peptide property prediction that supports both amino acid sequence inputs and SMILES representations, including peptides containing unnatural amino acids. By building on pretrained protein (ESM-2) and chemical (PeptideCLM) language models (Lin et al., 2023; Feller and Wilke, 2025a), PeptiVerse focuses on training lightweight predictor heads rather than full representation models, enabling efficient, scalable, and easily deployable property evaluation. These results highlight the heterogeneity of peptide property landscapes and motivate a flexible, property-aware modeling strategy rather than reliance on a single predictor.

Compared with prior SMILES-based peptide models such as PepLand (Zhang et al., 2025a), which rely on complex graph constructions and full-model retraining, PeptiVerse shows that foundation model embeddings paired with simple, well-regularized classifiers are sufficient (and often superior) for practical peptide property prediction. This design reduces computational cost while improving generalizability, making PeptiVerse compatible with modern peptide design pipelines. The advantage is particularly clear for binding affinity, where structure prediction confidence alone fails to track peptide-protein binding strength (Figure S4), whereas PeptiVerse yields strong, statistically significant agreement with experimental measurements (Table S6). Together, these results motivate fast, data-driven affinity predictors that complement rather than replace structural modeling, and position PeptiVerse as an open, extensible benchmark for therapeutic peptide discovery.

Importantly, PeptiVerse is designed as a lightweight and accessible framework that supports direct incorporation of property predictors into iterative and high-throughput design workflows, rather than limiting use to web-based post hoc evaluation (Tan et al., 2026). This design enables a natural application to generative peptide sequence modeling, where efficient property evaluation is required to guide optimization. In this setting, PeptiVerse predictors enable rapid assessment of therapeutically relevant properties, allowing generative models to produce sequences with a higher likelihood of translating in wet-lab settings (Tang et al., 2025c; Chen et al., 2025b;c; Tang et al., 2025d;b). Such guidance relies on fast reward evaluation to support gradient-based or iterative refinement over generated sequences (Nisonoff et al., 2025), particularly in multi-objective settings where properties such as binding affinity, solubility, and toxicity must be jointly optimized. In practice, PeptiVerse predictors have already been used to guide generative frameworks including PepTune (Tang et al., 2025c), which applies Monte Carlo Tree Guidance for multi-objective peptide optimization, TR2-D2 (Tang et al., 2025d), which performs trajectory-aware fine-tuning of peptide diffusion models, and moPPIt (Chen et al., 2026), which enables motif-specific peptide generation via discrete flow matching and is experimentally verified across multiple target classes. Together, these examples position PeptiVerse as a modular reward evaluation layer that integrates naturally into modern generative pipelines, supporting both inference-time steering and training-time optimization.

To ensure continued relevance, PeptiVerse will be updated regularly as new data and predictors become available. While several property models already achieve strong performance, others (i.e., chemically-modified peptide half-life in SMILES space) remain constrained by limited and heterogeneous experimental data. Importantly, the modular design of PeptiVerse enables seamless integration of improved representation models as they emerge; for example, retraining with newer embeddings such as ESM-C (EvolutionaryScale, 2024) provides performance

gains over ESM-2 on properties such as half-life and solubility (Tables S9 and S11), further supporting that advances in representation learning directly translate to improved predictive accuracy within this framework. We therefore encourage open-source deposition of peptide property measurements from both academic and industry. In parallel, we are developing new specialized binding predictors, including peptide isoform-specificity (Vincoff et al., 2025), motif-specificity (Chen et al., 2026), and metal-binding propensity (Zhang et al., 2025b), which will be incorporated as additional properties within the PeptiVerse framework. By providing a standardized, extensible, and openly accessible platform for integrating new data and models, PeptiVerse establishes a practical mechanism for expanding predictive scope, enabling property-guided generative design, and supporting the translation of next-generation peptide therapeutics.

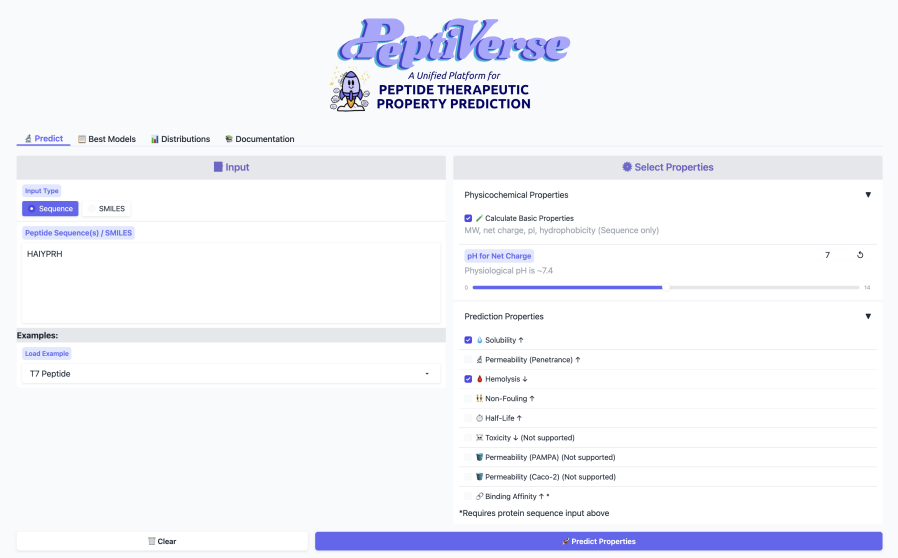

Figure 4: **Web interface of PeptiVerse hosted on HuggingFace.** The interface allows users to submit peptide inputs as either amino acid sequences or SMILES strings and select from a range of physicochemical and predictive properties for evaluation. Property options are dynamically enabled based on input type, and basic properties such as molecular weight, charge, and hydrophobicity can be computed alongside learned predictors. This design enables rapid, no-code peptide property assessment suitable for both exploratory analysis and integration into iterative design workflows.

Declarations

Data and Code Availability The full set of predictors and datasets is available through a simple API at https://huggingface.co/ChatterjeeLab/PeptiVerse. Code for plotting in R is available at https://github.com/ynuozhang/Peptiverse_R.git. For users who prefer a no-code interface, all predictors can also be easily run via an interactive HuggingFace Space at https://huggingface.co/spaces/ChatterjeeLab/PeptiVerse. All code is open source, with predictors and associated resources updated and maintained regularly.

References

Zhikai Zheng, Yao Zong, Yiyang Ma, Yucheng Tian, Yidan Pang, Changqing Zhang, and Junjie Gao. Glucagon-like peptide-1 receptor: mechanisms and advances in therapy. *Signal Transduction and Targeted Therapy*, 9(1), September 2024. ISSN 2059-3635. doi: 10.1038/s41392-024-01931-z. URL http://dx.doi.org/10.1038/s41392-024-01931-z.

Daniel J Drucker. Discovery of glp-1–based drugs for the treatment of obesity. *New England Journal of Medicine*, 392(6):612–615, 2025.

Leo Tianlai Chen, Zachary Quinn, Madeleine Dumas, Christina Peng, Lauren Hong, Moises Lopez-Gonzalez, Alexander Mestre, Rio Watson, Sophia Vincoff, Lin Zhao, et al. Target sequence-conditioned design of peptide binders using masked language modeling. *Nature Biotechnology*, pages 1–9, 2025a.

Lei Wang, Nanxi Wang, Wenping Zhang, Xurui Cheng, Zhibin Yan, Gang Shao, Xi Wang, Rui Wang, and Caiyun Fu. Therapeutic peptides: current applications and future directions. *Signal transduction and targeted therapy*, 7(1):48, 2022.

Sophia Tang, Emily L Han, and Michael J Mitchell. Peptide-functionalized nanoparticles for brain-targeted therapeutics. *Drug Delivery and Translational Research*, pages 1–20, 2025a.

Tianlai Chen, Lauren Hong, Vivian Yudistyra, Sophia Vincoff, and Pranam Chatterjee. Generative design of therapeutics that bind and modulate protein states. *Current Opinion in Biomedical Engineering*, 28:100496, 2023.

Kateřina Storchmannová, Martin Balouch, Jakub Juračka, František Štěpánek, and Karel Berka. Meta-analysis of permeability literature data shows possibilities and limitations of popular methods. *Molecular Pharmaceutics*, 22(3):1293–1304, February 2025. ISSN 1543-8392. doi: 10.1021/acs.molpharmaceut.4c00975. URL http://dx.doi.org/10.1021/acs.molpharmaceut.4c00975.

Guanyu Chen, Weirong Kang, Wanqiong Li, Shaomeng Chen, and Yanfeng Gao. Oral delivery of protein and peptide drugs: From non-specific formulation approaches to intestinal cell targeting strategies. *Theranostics*, 12(3):1419, 2022.

M. Werle and A. Bernkop-Schnürch. Strategies to improve plasma half life time of peptide and protein drugs. *Amino Acids*, 30(4):351–367, April 2006. ISSN 1438-2199. doi: 10.1007/s00726-005-0289-3. URL http://dx.doi.org/10.1007/s00726-005-0289-3.

Luciana Malavolta, Marcelo RS Pinto, Jamile H Cuvero, and Clóvis R Nakaie. Interpretation of the dissolution of insoluble peptide sequences based on the acid-base properties of the solvent. *Protein science*, 15(6):1476–1488, 2006.

Karolina L. Zapadka, Frederik J. Becher, A. L. Gomes dos Santos, and Sophie E. Jackson. Factors affecting the physical stability (aggregation) of peptide therapeutics. *Interface Focus*, 7(6):20170030, October 2017. ISSN 2042-8901. doi: 10.1098/rsfs.2017.0030. URL http://dx.doi.org/10.1098/rsfs.2017.0030.

Robert W. Kellermeyer. Hemolytic effect of therapeutic drugs: Clinical considerations of the primaquine-type hemolysis. *JAMA*, 180(5):388, May 1962. ISSN 0098-7484. doi: 10.1001/jama.1962.03050180034008a. URL http://dx.doi.org/10.1001/jama.1962.03050180034008a.

Patrick Brendan Timmons and Chandralal M Hewage. Happenn is a novel tool for hemolytic activity prediction for therapeutic peptides which employs neural networks. *Scientific reports*, 10(1):10869, 2020.

Shaoyi Jiang and Zhiqiang Cao. Ultralow-fouling, functionalizable, and hydrolyzable zwitterionic materials and their derivatives for biological applications. *Advanced Materials*, 22(9):920–932, February 2010. ISSN 1521-4095. doi: 10.1002/adma.200901407. URL http://dx.doi.org/10.1002/adma.200901407.

Chakradhar Guntuboina, Adrita Das, Parisa Mollaei, Seongwon Kim, and Amir Barati Farimani. Peptidebert: A language model based on transformers for peptide property prediction. *The Journal of Physical Chemistry Letters*, 14(46):10427–10434, 2023.

Jianan Li, Keisuke Yanagisawa, Masatake Sugita, Takuya Fujie, Masahito Ohue, and Yutaka Akiyama. Cycpeptmpdb: a comprehensive database of membrane permeability of cyclic peptides. *Journal of Chemical Information and Modeling*, 63(7):2240–2250, 2023.

Justine N DeGruyter, Lara R Malins, and Phil S Baran. Residue-specific peptide modification: a chemist's guide. *Biochemistry*, 56(30):3863–3873, 2017.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood Van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36:64331–64379, 2023.

Jiaqing Xie, Yuqiang Li, and Tianfan Fu. Deepprotein: deep learning library and benchmark for protein sequence learning. *Bioinformatics*, 41(10):btaf165, 2025.

Gabriel Cabas-Mora, Anamaría Daza, Nicole Soto-García, Valentina Garrido, Diego Alvarez, Marcelo Navarrete, Lindybeth Sarmiento-Varón, Julieta H Sepúlveda Yañez, Mehdi D Davari, Frederic Cadet, et al. Peptipedia v2. 0: A peptide sequence database and user-friendly web platform. a major update. *Database*, 2024:baae113, 2024.

Li Fu, Shaohua Shi, Jiacai Yi, Ningning Wang, Yuanhang He, Zhenxing Wu, Jinfu Peng, Youchao Deng, Wenxuan Wang, Chengkun Wu, et al. Admetlab 3.0: an updated comprehensive online admet prediction platform enhanced with broader coverage, improved performance, api functionality and decision support. *Nucleic acids research*, 52(W1): W422–W431, 2024.

Kyle Swanson, Parker Walther, Jeremy Leitz, Souhrid Mukherjee, Joseph C Wu, Rabindra V Shivnaraine, and James Zou. Admet-ai: a machine learning admet platform for evaluation of large-scale chemical libraries. *Bioinformatics*, 40(7):btae416, 2024.

Ruochi Zhang, Haoran Wu, Chang Liu, Qian Yang, Yuting Xiu, Kewei Li, Ningning Chen, Yu Wang, Yan Wang, Xin Gao, et al. Pepland: a large-scale pre-trained peptide representation model for a comprehensive landscape of both canonical and non-canonical amino acids. *Briefings in bioinformatics*, 26(4):bbaf367, 2025a.

Leyao Wang, Rishab Pulugurta, Pranay Vure, Yinuo Zhang, Aastha Pal, and Pranam Chatterjee. Pepdora: A unified peptide language model via weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2410.20667*, 2024.

Mehrad Ansari and Andrew D. White. Serverless prediction of peptide properties with recurrent neural networks. *Journal of Chemical Information and Modeling*, 63(8):2546–2553, April 2023. doi: 10.1021/acs.jcim.2c01317. URL https://doi.org/10.1021/acs.jcim.2c01317.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Aaron L Feller and Claus O Wilke. Peptide-aware chemical language model successfully predicts membrane diffusion of cyclic peptides. *Journal of Chemical Information and Modeling*, 65(2):571–579, 2025a.

Tong Chen, Yinuo Zhang, Sophia Tang, and Pranam Chatterjee. Multi-objective-guided discrete flow matching for controllable biological sequence design. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*, 2025b. URL https://openreview.net/forum?id=8YIMLoHP9J.

Tong Chen, Yinuo Zhang, and Pranam Chatterjee. Areuredi: Annealed rectified updates for refining discrete flows with multi-objective guidance. *arXiv preprint arXiv:2510.00352*, 2025c.

Shrey Goel, Peregrine Michael Schray, Yinuo Zhang, Sophia Vincoff, Huong T. Kratochvil, and Pranam Chatterjee. Token-level guided discrete diffusion for membrane protein design. In *NeurIPS AI4Science Workshop*, 2025. URL https://openreview.net/forum?id=I0hzddNny7.

Sophia Tang, Yinuo Zhang, Alexander Tong, and Pranam Chatterjee. Gumbel-softmax flow matching with straight-through guidance for controllable biological sequence generation. *arXiv preprint arXiv:2503.17361*, 2025b.

Sophia Tang, Yinuo Zhang, and Pranam Chatterjee. Peptune: De novo generation of therapeutic peptides with multi-objective-guided discrete diffusion. In *Forty-second International Conference on Machine Learning*, 2025c. URL https://openreview.net/forum?id=FQoy1Y1Hd8.

Sophia Tang, Yuchen Zhu, Molei Tao, and Pranam Chatterjee. Tr2-d2: Tree search guided trajectory-aware fine-tuning for discrete diffusion. *arXiv preprint arXiv:2509.25171*, 2025d.

Yinuo Zhang, Divya Srijay, Zachary Quinn, and Pranam Chatterjee. Metalorian: De novo generation of heavy metal-binding peptides with classifier-guided diffusion sampling. *bioRxiv*, pages 2025–07, 2025b.

Sophia Vincoff, Oscar Davis, Ismail Ilkan Ceylan, Alexander Tong, Joey Bose, and Pranam Chatterjee. SOAPIA: Siamese-guided generation of off target-avoiding protein interactions with high target affinity. In *ICML 2025 Workshop on Scaling Up Intervention Models*, 2025. URL https://openreview.net/forum?id=j0OpIG7leX.

Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

Valeria Simoncini. Variable accuracy of matrix-vector products in projection methods for eigencomputation. *SIAM journal on numerical analysis*, 43(3):1155–1174, 2005.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

Jasper De Landsheere, Anton Zamyatin, Johannes Karwounopoulos, and Esther Heid. Chemtorch: A deep learning framework for benchmarking and developing chemical reaction property prediction models. 2025.

Malak Pirtskhalava, Anthony A Amstrong, Maia Grigolava, Mindia Chubinidze, Evgenia Alimbarashvili, Boris Vishnepolsky, Andrei Gabrielian, Alex Rosenthal, Darrell E Hurt, and Michael Tartakovsky. Dbaasp v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic acids research*, 49(D1):D288–D297, 2021.

Anand Singh Rathore, Shubham Choudhury, Akanksha Arora, Purva Tijare, and Gajendra PS Raghava. Toxinpred 3.0: An improved method for predicting the toxicity of peptides. *Computers in biology and medicine*, 179:108926, 2024.

Pawel Smialowski, Gero Doose, Phillipp Torkler, Stefanie Kaufmann, and Dmitrij Frishman. Proso ii–a new method for protein solubility prediction. *The FEBS journal*, 279(12): 2192–2200, 2012.

Shipra Jain, Srijanee Gupta, Sumeet Patiyal, and Gajendra PS Raghava. Thpdb2: compilation of fda approved therapeutic peptides and proteins. *Drug Discovery Today*, 29(7): 104047, 2024.

Deepika Mathur, Satya Prakash, Priya Anand, Harpreet Kaur, Piyush Agrawal, Ayesha Mehta, Rajesh Kumar, Sandeep Singh, and Gajendra PS Raghava. Peplife: a repository of the half-life of peptides. *Scientific reports*, 6(1):36617, 2016.

Vera D'Aloisio, Paolo Dognini, Gillian A Hutcheon, and Christopher R Coxon. Peptherdia: database and structural composition analysis of approved peptide therapeutics and diagnostics. *Drug Discovery Today*, 26(6):1409–1419, 2021.

Rainier Barrett, Shaoyi Jiang, and Andrew D White. Classifying antimicrobial and multifunctional peptides with bayesian network models. *Peptide Science*, 110(4):e24079, 2018.

Jan Veleckỳ, Marie Hamsikova, Jan Stourac, Milos Musil, Jiri Damborsky, David Bednar, and Stanislav Mazurenko. Soluprotmutdb: A manually curated database of protein solubility changes upon mutations. *Computational and Structural Biotechnology Journal*, 20:6339–6347, 2022.

Alex Avdeef. The rise of pampa. *Expert opinion on drug metabolism & toxicology*, 1(2): 325–342, 2005.

Per Artursson, Katrin Palm, and Kristina Luthman. Caco-2 monolayers in experimental and theoretical predictions of drug transport. *Advanced drug delivery reviews*, 64:280–289, 2012.

Yipin Lei, Shuya Li, Ziyi Liu, Fangping Wan, Tingzhong Tian, Shao Li, Dan Zhao, and Jianyang Zeng. A deep-learning framework for multi-level peptide–protein interaction prediction. *Nature communications*, 12(1):5465, 2021.

Wensi Zhu, Aditi Shenoy, Petras Kundrotas, and Arne Elofsson. Evaluation of alphafold-multimer prediction on multi-chain protein complexes. *Bioinformatics*, 39(7):btad424, 2023.

Chunxiang Peng, Wentao Ni, Quancheng Liu, Gang Hu, and Wei Zheng. A comprehensive benchmarking of the alphafold3 for predicting biomacromolecules and their interactions. *Briefings in Bioinformatics*, 26(6):bbaf616, 2025.

The OpenFold3 Team. Openfold3-preview, 2025. URL https://github.com/aqlaboratory/openfold-3.

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019.

Shashank Mohan Jain. Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*, pages 51–67. Springer, 2022.

Xiaorong Tan, Qianhui Liu, Mengting Zhou, Yanpeng Fang, Defang Ouyang, Wenbin Zeng, and Jie Dong. pepadmet: A novel computational platform for systematic admet evaluation of peptides. *Journal of Chemical Information and Modeling*, 2026.

Hunter Nisonoff, Junhao Xiong, Stephan Allenspach, and Jennifer Listgarten. Unlocking guidance for discrete state-space diffusion and flow models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=XsgHl54yO7.

Tong Chen, Zachary Quinn, Kunal Mishra, Erin C. O'Connor, Sophia E. Silver, Yinuo Zhang, Mary Jo Valencia, Ying Mei, Jacques Behmoaras, Leonardo M.R. Ferreira, and Pranam Chatterjee. moppit: De novo generation of motif-specific and functionally active peptide binders via discrete flow matching. *bioRxiv*, January 2026. doi: 10.1101/2024.07.31.606098. URL http://dx.doi.org/10.1101/2024.07.31.606098.

EvolutionaryScale. Esm cambrian: Revealing the mysteries of proteins with unsupervised learning, Dec 2024. URL https://www.evolutionaryscale.ai/blog/esm-cambrian.

Giorgio Ottaviani, Sophie Martel, and Pierre-Alain Carrupt. Parallel artificial membrane permeability assay: a new membrane for the fast prediction of passive human skin permeability. *Journal of medicinal chemistry*, 49(13):3948–3954, 2006.

Richard B Van Breemen and Yongmei Li. Caco-2 cell permeability assays to measure drug absorption. *Expert opinion on drug metabolism & toxicology*, 1(2):175–185, 2005.

Indrani Banerjee, Ravindra C Pangule, and Ravi S Kane. Antifouling coatings: recent developments in the design of surfaces that prevent fouling by proteins, bacteria, and marine organisms. *Advanced materials*, 23(6):690–718, 2011.

Aaron L Feller and Claus O Wilke. p2smi: A toolkit enabling smiles generation and property analysis for noncanonical and cyclized peptides. *Journal of Open Source Software*, 10(116): 8319, 2025b.

Helen M Berman, John D Westbrook, Margaret J Gabanyi, Wendy Tao, Raship Shah, Andrei Kouranov, Torsten Schwede, Konstantin Arnold, Florian Kiefer, Lorenza Bordoli, et al. The protein structure initiative structural genomics knowledgebase. *Nucleic acids research*, 37(suppl_1):D365–D368, 2009.

Andrei Kouranov, Lei Xie, Joanna de la Cruz, Li Chen, John Westbrook, Philip E Bourne, and Helen M Berman. The rcsb pdb information portal for structural genomics. *Nucleic acids research*, 34(suppl_1):D302–D305, 2006.

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic acids research*, 52(D1):D1265–D1275, 2024.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *BioRxiv*, pages 2023–10, 2023.

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016): 493–500, 2024.

NVIDIA Corporation. Explore biology models | try NVIDIA NIM APIs, 2025. URL https://build.nvidia.com/explore/biology.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2019.

Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422, 2009.

Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, Peter Gedeck, Gareth Jones, Eisuke Kawashima, Dan Nealschneider, et al. rdkit/rdkit: 2025_03_1 (q1 2025) release. *Zenodo*, 2025.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

Miguel Carrasco, Julio López, and Sebastián Maldonado. Epsilon-nonparallel support vector regression. *Applied Intelligence*, 49(12):4223–4236, 2019.

Sebastian Raschka, Joshua Patterson, and Corey Nolet. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *arXiv preprint arXiv:2002.04803*, 2020.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12: 2825–2830, 2011.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Vianney Taquet, Vincent Blot, Thomas Morzadec, Louis Lacombe, and Nicolas Brunel. Mapie: an open-source library for distribution-free uncertainty quantification. *arXiv preprint arXiv:2207.12274*, 2022.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

HU Mei, Zhi H Liao, Yuan Zhou, and Shengshi Z Li. A new set of amino acid descriptors and its application in peptide qsars. *Peptide Science: Original Research on Biomolecules*, 80(6):775–786, 2005.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

S1 Supplementary Methods

S1.1 Data Collection and Preparation

Throughout this work, we distinguish peptide inputs by representation modality rather than biological origin. "Amino acid" inputs refer to canonical sequence-based representations processed by protein language models, while "SMILES" inputs refer to chemistry-aware molecular representations, which may include both canonical and non-canonical peptides.

Hemolysis Hemolysis data were retrieved from PeptideBert and peptide-dashboard (Guntuboina et al., 2023; Ansari and White, 2023), and cross-validated against the original experimental records in DBAASP v3.0 (Pirtskhalava et al., 2021). Peptides labeled as 1 are considered hemolytic, whereas 0 denotes non-hemolytic activity. The final dataset comprised 4,765 hemolytic and 1,311 non-hemolytic peptide entries.

Permeability All permeability annotations were obtained from PepLand (Zhang et al., 2025a), which sources experimental measurements from CycPeptMPDB (Li et al., 2023). The *noncanonical* dataset contains 7,334 non-canonical peptides with reported permeability values measured using either PAMPA (Avdeef, 2005) or Caco-2 assays (Artursson et al., 2012). Permeability is reported as $\log P_{\text{exp}}$, the logarithm of the effective permeability coefficient, which reflects the peptide's lipophilicity and its ability to passively diffuse across lipid membranes.

PAMPA and Caco-2 assays quantify different biological processes, with PAMPA measuring passive membrane permeability and Caco-2 assays relating more closely to intestinal absorption potential (Avdeef, 2005; Ottaviani et al., 2006; Van Breemen and Li, 2005). For this reason, the two assay types were handled separately during data preparation, not following PepLand's combined training strategy. There were 6,869 PAMPA sequences and 606 Caco-2 sequences collected in the end. Following CycPeptMPDB conventions, peptides with $\log P_{\text{exp}} \geq -6.0$ were labeled as high permeability, and the remaining peptides as weak permeability. (Li et al., 2023). The predictive tasks associated with these datasets are called Permeability (PAMPA) and Permeability (Caco-2).

The canonical permeability dataset contains 1,162 cell-penetrating peptides and 1,162 non-penetrating peptides, curated from PepLand (Zhang et al., 2025a). This dataset was constructed to balance peptide length distributions between positive and negative classes. Positive peptides were originally collected from 22 independent cell-penetrating peptide studies, while negative examples were sourced from UniProt. A notable fraction of the positive peptides (approximately 8.7%) exhibited low sequence complexity, defined as a ratio of peptide length to the number of unique amino acids greater than 5, compared to only 0.95% among non-penetrating peptides. Although these low-complexity sequences are likely engineered, they were retained, as they may encode informative features relevant to membrane permeability (Zhang et al., 2025a). The predictive task associated with this dataset is called Permeability_CPP.

Non-Fouling Non-fouling annotations were obtained from PeptideBERT and Peptide-Dashboard (Guntuboina et al., 2023; Ansari and White, 2023), both of which source data from the dataset of Barrett et al. (2018). In this context, non-fouling peptides are defined as sequences that resist nonspecific protein adsorption, while fouling peptides permit such adsorption, which can lead to functional loss and reduced performance (Barrett et al., 2018; Banerjee et al., 2011). Peptides labeled as 1 are classified as non-fouling, and those labeled as 0 are considered fouling. The curated dataset comprised 13,580 non-fouling and 3,600 fouling peptide entries.

Toxicity Toxicity data were obtained from ToxinPred3.0, which provides canonical amino acid peptide sequences with experimentally validated toxicity labels (Rathore et al., 2024). The dataset contains 5,518 toxic and 5,518 non-toxic peptides, where label 1 denotes toxic and 0 denotes non-toxic. Amino acid sequences were converted into SMILES format with `fasta2smi` command from `p2smi` (Feller and Wilke, 2025b). Molecular redundancy was reduced by clustering peptides using Morgan fingerprints (radius 2, 2048 bits, including chirality) and RDKit's Butina clustering algorithm with a similarity threshold of 0.6.

Solubility Solubility labels follow the protocol established in PROSO II (Smialowski et al., 2012), as used in PeptideBERT (Guntuboina et al., 2023), with additional sequences incorporated from SoluProtMutDB (Velecký et al., 2022). In PROSO II, the soluble class (1) was assigned based on experimental metadata recorded in pepcDB, following the Protein Structure Initiative (PSI) pipelines (Berman et al., 2009). A sequence was labeled soluble once it reached the "Soluble" stage or any downstream stage. Additional soluble entries were derived from PDB records (up to 2010) annotated with expression from *E. coli*, as crystallographic and NMR structure determination requires a protein that has been purified in solution. The insoluble class (0) consisted of constructs that remained in the "not Soluble" state for at least eight months, based on the different pepcDB releases. This yields a combined dataset of 8,785 soluble and 9,668 negative sequences in total.

Binding Affinity Binding affinity prediction was performed using paired protein sequences and peptide SMILES from the PepLand dataset (Zhang et al., 2025a), which aggregates different experimental measurements together (K_d, K_i, IC_{50}). Their data collection protocol follows CAMP (Lei et al., 2021), which subsets data from RCSB PDB (Kouranov et al., 2006) and DrugBank entries containing the label of "peptide" (Knox et al., 2024). All scores were transformed based on the negative logarithm of the original affinity data into a unified scale. A shared unit scale was roughly designed to indicate the strength of binding: with 9 indicating strong nM to pM binders; 7-9 indicating nM to μM medium binders, and < 7 indicating weak μM binders. The data split was based on affinity score distribution matching, ensuring both splits containing the similar distribution style of data. The final dataset contained 1,436 peptide–protein pairs with canonical peptide sequences (amino acid inputs) and 1,597 pairs with SMILES-based peptides containing noncanonical chemistry.

$$\text{affinity\_score} = -\log_{10}(K_{d/i} \text{ in M}) \tag{1}$$

Thus, higher values indicate stronger binding, which is more intuitive for modeling.

Stability Stability data for amino acid sequence inputs were collected and organized from TAPE (Rao et al., 2019) and SaProt (Su et al., 2023) (Huggingface: `SaProtHub/Dataset-Stability-TAPE`) were used to pretrain stability predictors before half-life modeling. Stability was treated as a continuous score reflecting the ability of proteins to remain folded above concentration thresholds. A total of 68,845 sequences were collected.

Half-Life Half-life data were compiled from THPdb2 (Jain et al., 2024), PEPlife (Mathur et al., 2016), and PepTherDia (D'Aloisio et al., 2021). Only human serum measurements were retained. Reported values across datasets vary in units and granularity. All half-life measurements were therefore converted into hours for interpretability. Details of the data construction are provided in the Supplementary Information. The curated dataset comprised 130 amino acid sequence entries and 245 SMILES-based sequences.

Protein-peptide ipTM scores The interface predicted TM-score (ipTM) is a confidence metric originally designed to assess the accuracy of predicted protein–protein interaction interfaces in multimeric structure prediction models (Abramson et al., 2024). All ipTM scores for protein-peptide pairs, including SMILES represented pairs, were retrieved from API calling for the OpenFold3 (The OpenFold3 Team, 2025) through NVIDIA NGC platform (NVIDIA Corporation, 2025). MSA profiles for the target proteins were constructed via MMseqs2 (Steinegger and Söding, 2017) against the Uniref30 database (Consortium, 2019).

Additional Properties Additional physicochemical features, including isoelectric point, molecular weight, and hydrophobicity, were computed using utilities from `Biopython` package (Cock et al., 2009). These properties were calculated dynamically with a tunable pH parameter to account for protonation-state dependence.

Data Splitting All properties defined on amino acid–based peptide representations were split using a shared clustering strategy. Sequences from all amino acid sequence datasets were clustered independently using MMseqs2 (Steinegger and Söding, 2017) with identical parameters (`-min-seq-id 0.3 -c 0.8 -cov-mode 0`) and split using an 80/20 cluster-level split to prevent sequence leakage between training and evaluation subsets. Amino acid sequences were converted into SMILES format with `fasta2smi` command from `p2smi` (Feller and Wilke, 2025b). For properties evaluated using both sequence and SMILES

Table S1: **Dataset composition across peptide property prediction tasks.** Classification tasks report the number of samples per class (0/1), where 0 is negative, and 1 is positive. Regression tasks report the total number of samples (N). Counts are shown separately for amino acid and SMILES sequence representations, illustrating differences in data availability and class balance across properties.

Properties	Amino Acid Sequences		SMILES Sequences	
	0	1	0	1
Classification				
Hemolysis	4765	1311	4765	1311
Non-Fouling	13580	3600	13580	3600
Solubility	9668	8785	-	-
Permeability_CPP	1162	1162	-	-
Toxicity	-	-	5518	5518
Regression (N)				
Permeability (PAMPA)	-	-	6869	
Permeability (CACO2)	-	-	606	
Half-Life	130		245	
Binding Affinity	1433		1702	

inputs (hemolysis, non-fouling, solubility, permeability_cpp), the original cluster-based train/validation assignments were preserved after conversion, ensuring fair comparison across input modalities. For datasets natively represented in SMILES space, molecules were clustered using Morgan fingerprints and Tanimoto distance via RDKit (Landrum et al., 2025), followed by an 80/20 cluster-level split. Binding affinity prediction, which involves paired peptide–protein inputs across two molecular modalities, was instead split by matching affinity score distributions between training and evaluation sets. Detailed dataset sizes, label distributions, and train/validation splits for all properties are summarized in Table S1 and Figure 2.

S1.2 Model Architecture and Training

Sequence and SMILES Representations Protein sequences were represented using embeddings from ESM-2 (`esm2_t33_650M_UR50D`) (Lin et al., 2023) while we embedded peptide sequences using either ESM2, PeptideCLM (`PeptideCLM-23M-all`) (Feller and Wilke, 2025a), which employs a tokenizer of size 586 designed to capture noncanonical peptide chemistry, or ChemBERTA (`ChemBERTa-77M-MLM`) (Chithrananda et al., 2020), trained across millions of chemical structures curated from PubChem. See Supplementary Methods for details on other embedding approaches including ESM-C and baseline embeddings. Depending on the downstream model, embeddings were either average-pooled across sequence positions to form fixed-length representations or retained in unpooled, position-resolved form to preserve residue-level and positional information. This distinction allowed different model classes to exploit either global sequence summaries or fine-grained positional structure.

Predictor Architectures For properties including toxicity, solubility, hemolysis, non-fouling, and permeability, lightweight predictor heads were trained on frozen foundational embeddings. We evaluated a broad set of predictor architectures, including multilayer perceptrons (MLPs), convolutional neural networks (CNNs), and transformer-based models, as well as classical statistical learners such as Elastic Nets (ENET) (Zou and Hastie, 2005), support vector machines (SVMs) (Hearst et al., 1998), epsilon-support vector regression (SVR) (Carrasco et al., 2019), and XGBoost (XGB) (Chen et al., 2015). ENET and SVM classifiers were implemented using the `RAPIDS cuML` library (Raschka et al., 2020) to enable GPU acceleration, while SVR models were implemented using `scikit-learn` (Pedregosa et al., 2011). Classification models were trained using standard objective functions appropriate to each method, whereas regression models optimized mean squared error. All models were trained using their canonical formulations without architectural modification. For CNN- and transformer-based predictors, unpooled embeddings were used to allow convolutional kernels and self-attention mechanisms to explicitly exploit positional structure, while pooled embeddings were used for MLP and tree-based models operating on fixed-dimensional inputs.

Hyperparameter Optimization Hyperparameters for all models were optimized using the Optuna framework (Akiba et al., 2019), with an initial target of 200 optimization trials per configuration (Supplementary Tables S13–S16). For parameters sampled on a logarithmic

scale, values were drawn uniformly in log space to efficiently explore multiple orders of magnitude. For regression tasks, both mean squared error (MSE) and Huber loss were evaluated, using the latter to improve robustness to outliers via its δ parameter controlling the transition between L1 and L2 penalties. For computationally intensive configurations exceeding approximately 10 hours of wall-clock training time, the number of Optuna trials was reduced to 50 or 20 while preserving representative coverage of the hyperparameter space. Results of the hyperparameter exploration are reported in Supplementary Figures S2 and S5.

Binding Affinity Prediction Binding affinity prediction was formulated using a transformer-based architecture with cross multi-head attention to learn a joint latent representation between peptide and protein modalities. The model architecture was fixed across experiments, with variation introduced only through peptide embedding initialization. Peptide inputs were represented using either ESM-2 or PeptideCLM embeddings (Lin et al., 2023; Feller and Wilke, 2025a), in pooled or unpooled form, while protein targets were consistently represented using ESM-2 embeddings. The model produced a continuous affinity score, which was additionally mapped to three discrete affinity classes, yielding a multitask formulation aligned with other property predictors. Hyperparameters were optimized using 200 Optuna trials, and Spearman's rank correlation coefficient (ρ) was used as the primary model selection criterion (Supplementary Table S17; Supplementary Figure S2).

Half-Life Prediction Peptide half-life prediction was formulated as a regression task using both amino acid sequence inputs and SMILES-based representations. For amino acid–based inputs, models were initialized from pre-trained stability predictors trained on the wild-type stability dataset and subsequently fine-tuned on half-life data. The same model families used for binary property prediction were retained. Four regression configurations were evaluated: (i) XGBoost predicting half-life in hours, (ii) XGBoost predicting $\log(1 + \text{half-life})$, (iii) a transformer-based regressor predicting half-life in hours, and (iv) a transformer-based regressor predicting $\log(1 + \text{half-life})$. Hyperparameters were optimized using Optuna with five-fold cross-validation, using 200 trials for XGB, SVR, and ENET models and 50 trials for CNN, MLP, and Transformer models, with 20 training epochs per trial.

The transformation equation is defined as follows:

$$\text{half-life\_score} = \ln(1 + t_{1/2}) \tag{2}$$

where $t_{1/2}$ is the half_life in hours.

For SMILES-based half-life prediction, models were trained directly on peptide SMILES embeddings without intermediate stability pretraining. Predictor architectures matched those used for amino acid–based half-life prediction and were trained to predict $\log(1 + \text{half-life})$. Evaluation was performed using five-fold cross-validation. Optuna optimization employed 200 trials for XGB, SVR, and ENET models and 50 trials for CNN, MLP, and Transformer models, with transformer-based models trained for 100 epochs per trial.

Evaluation Metrics Model evaluation metrics were selected according to task type. Classification performance was assessed using F1 score and area under the receiver operating characteristic curve (AUC) with results reported in Supplementary Figure S1 and Supplementary Table S4. Regression performance was evaluated using Pearson correlation coefficient (r), Spearman's rank correlation coefficient (ρ), and the coefficient of determination (R^2), with results summarized in Supplementary Figure S3.

Model Stability, Confidence Intervals, and Uncertainty Estimation All deep learning models (MLP, CNN, Transformer) were retrained five times with distinct random seeds (1986, 42, 0, 123, 12345) using fixed optimal hyperparameters from OPTUNA. The 95% confidence intervals for classification (F1, AUC) and regression (Spearman ρ, RMSE, R^2) metrics were estimated as $\bar{x} \pm t_{0.025,4} \cdot \frac{s}{\sqrt{5}}$. Per-sequence uncertainty was quantified as the standard deviation of predictions across seeds.

For classical ML models (SVM, ElasticNet), optimization is convex and therefore deterministic given fixed data and hyperparameters (Boyd and Vandenberghe, 2004), leading to negligible variability across random seeds. For XGBoost, while stochastic subsampling of features and instances introduces minor seed dependence (Chen et al., 2015), empirical variability is

small relative to data sampling variability. Confidence intervals for all three model types were therefore derived via non-parametric percentile bootstrap ($n = 2000$) over held-out validation predictions $(y_i, \hat{y}_i)$ (Efron and Tibshirani, 1994) without retraining. The percentile method was preferred over symmetric t-intervals to respect the bounded range of Spearman $\rho \in [-1, 1]$, with Fisher z-transform intervals computed as a cross-check when $\rho > 0.9$. The decision threshold was fixed on the full validation set and held constant across bootstrap resamples to avoid optimistic bias. Per-sequence uncertainty was estimated as the probability margin from the decision boundary for probabilistic classifiers ($c_i = 2|p_i - 0.5|$), prediction variance across boosting steps for XGBoost, and empirical residual intervals ($\hat{y}_i \pm z_{0.975}\hat{\sigma}$) for deterministic regression models (SVR, ElasticNet).

For the half-life regression models, 95% confidence intervals were computed by non-parametric percentile bootstrap over out-of-fold prediction pairs $(y_i, \hat{y}_i)$ obtained from cross-validation without retraining. For each bootstrap resample, RMSE, MAE, R^2, and Spearman's ρ were recomputed, and the 2.5th and 97.5th percentiles of the resulting empirical distributions were reported as the interval bounds. In addition to the point estimate on the full validation set, we also report the bootstrap mean and standard deviation for each metric.

Ablation of Full Fine-Tuning and LoRA for Half-Life Prediction To further assess the role of parameter-efficient adaptation in this data-limited setting, we performed an additional half-life transfer-learning ablation comparing full fine-tuning and low-rank adaptation (LoRA) using the stability-pretrained transformer (Hu et al., 2022) (Table S12). This comparison was restricted to the transformer architecture, as it was the strongest sequence-based stability model and the most appropriate setting for LoRA-based adaptation. In the LoRA setting, the pretrained stability checkpoint was first loaded, after which selected linear layers in the transformer were replaced with LoRA-augmented modules while the backbone weights were frozen; only the low-rank update matrices and the final regression head were trained. Performance was evaluated for both raw and log-transformed half-life targets. Full fine-tuning outperformed LoRA under both target formulations, with the best results obtained for full fine-tuning on log-transformed half-life, whereas LoRA remained competitive but consistently lower-performing. These results suggest that transfer from stability prediction to half-life prediction benefits from broader adaptation of the pretrained representation, while log-space prediction provides a more favorable learning target for this task.

Inference-Time Confidence Estimation To provide confidence scores per-prediction at inference time, different strategies were applied according to the model class, following the framework of (Lakshminarayanan et al., 2017).

For classification tasks, uncertainty is quantified as the predictive entropy of the model's output probability. For deep learning classifiers (MLP, CNN, Transformer), inference-time uncertainty was quantified using the deep ensemble of five independently seeded models. Given a query input $\mathbf{x}$, each model $m \in \{1, \ldots, 5\}$ produces output probability $p_{\theta_m}(y|\mathbf{x})$. The ensemble predictive distribution is the uniformly-weighted mixture $\bar{p}(y|\mathbf{x}) = \frac{1}{5}\sum_{m=1}^{5} p_{\theta_m}(y|\mathbf{x})$. For classical classifiers (XGBoost, SVM), $\bar{p}(y|\mathbf{x})$ is the single model's output probability. For linear classifiers (ElasticNet) that produce a decision function score $f(\mathbf{x})$ rather than a native probability, $\bar{p}(y|\mathbf{x}) = \sigma(f(\mathbf{x}))$ where σ denotes the sigmoid function. In both cases, per-sample uncertainty is quantified as the binary predictive entropy:

$$\mathcal{H}[\bar{p}] = -\bar{p}\log\bar{p} - (1-\bar{p})\log(1-\bar{p}) \tag{3}$$

High entropy indicates either ensemble disagreement (epistemic uncertainty) or collectively diffuse predictions (aleatoric uncertainty). The deep ensemble formulation is preferred over MC Dropout (Gal and Ghahramani, 2016) as deep ensembles explore multiple modes of the loss landscape via independent random initialization, whereas MC Dropout samples within a single basin, leading to systematically overconfident predictions on out-of-distribution inputs (Lakshminarayanan et al., 2017).

For regression tasks, uncertainty estimation depends on the model class. For deep learning regression models (MLP, CNN, Transformer), inference-time uncertainty is reported as the

standard deviation of point predictions across the five ensemble members:

$$\sigma_{\text{ens}}(\mathbf{x}) = \sqrt{\frac{1}{5}\sum_{m=1}^{5}\left(\hat{y}_m(\mathbf{x}) - \bar{y}(\mathbf{x})\right)^2} \tag{4}$$

where $\bar{y}(\mathbf{x}) = \frac{1}{5}\sum_m \hat{y}_m(\mathbf{x})$ is the ensemble mean prediction. For deterministic classical regression models (SVR, ElasticNet), for which seed-based ensembling is inapplicable due to convex optimization yielding identical solutions, and for XGBoost, where retraining seed ensembles was outside the scope of this work, inference-time prediction intervals were constructed using conformal prediction via the MAPIE framework (Taquet et al., 2022). We implement the residual normalized conformity score (Taquet et al., 2022; Lei et al., 2018) using split conformal prediction. Direct use of MAPIE's sklearn API was precluded by our pipeline architecture: base models take raw sequences as input and embed them via GPU-accelerated transformer forward passes (ESM2-650M, ChemBERTa), and the binding affinity model accepts dual inputs incompatible with sklearn's single-array `fit(X, y)` interface. We therefore implement the equivalent procedure manually: an auxiliary XGBoost $\hat{\sigma}$ is trained on held-out embeddings and absolute residuals, conformity scores $s_i = |y_i - \hat{y}_i|/\hat{\sigma}(\mathbf{x}_i)$ are computed, and the finite-sample corrected quantile q is taken at level $\lceil (n+1)(1-\alpha)\rceil/n$ (Lei et al., 2018).

S1.3 Extra Benchmarking Experiments with Different Embedding Schemes

One-Hot and VHSE embeddings To compare model performance on baseline representations, all wild-type amino acid predictors were retrained with one-hot and VHSE embeddings (Tables S7, S8, S9). These embeddings consist of fixed vectors for each amino acid regardless of their position within the sequence; unlike ESM-2 and other learned representations, they are not context dependent. All models were trained for 200 trials of Optuna hyperparameter search. One-hot representation assigns each amino acid to one index number and that number will count from 0 to 19, covering all 20 wild-type amino acids. VHSE representations (Mei et al., 2005) are precomputed vectors for each of the 20 amino acids, containing 8 numerical descriptors of hydrophobic, steric, and electronic properties.

ESM-C embeddings We further compare embeddings derived from ESM-C (`esmc-300m-2024-12`) (EvolutionaryScale, 2024), an extension of the ESM-2 models specifically trained for producing information-rich representations. We provide the code and weights we used to train models on ESM-C embeddings. Notably ESM-C model outperform ESM-2 embeddings on the solubility and half-life tasks, but not too much. Considering that ESM-C models require token access for faster embedding generation, we do not host this feature on our repository. Users will need to generate their own embeddings and run our models locally on their peptides.

ECFP embeddings The SMILES baseline training used ECFP embeddings: Extended-connectivity Fingerprints for SMILES strings (Rogers and Hahn, 2010). ECFP vectors are fixed-length representations of molecules that encode the presence of local molecular substructures, generated by iteratively hashing each atom's neighborhood out to a user-specified radius. Since there is no "unpooled" version of an ECFP vector, each deep neural network used the same vector as ML models.

Extra Structural Validations

For the validation sets in the binding affinity data, we collected ipTM scores from OpenFold3 (The OpenFold3 Team, 2025) and AlphaFold3 (Abramson et al., 2024). The poses predicted by OpenFold3 were used as inputs for AutoDock VINA docking (v1.1.2) (Trott and Olson, 2010). Fewer results are presented in Figure S4 for some methods, as structures exhibiting steric clashes or unfavorable interatomic distances were excluded.

S1.4 Supplementary Figures and Tables

Figure S1: **Per class performance across models.** The best performing models were selected based on validation performance, refit on the validation set, and evaluated for final reporting. Properties without data are not reported.

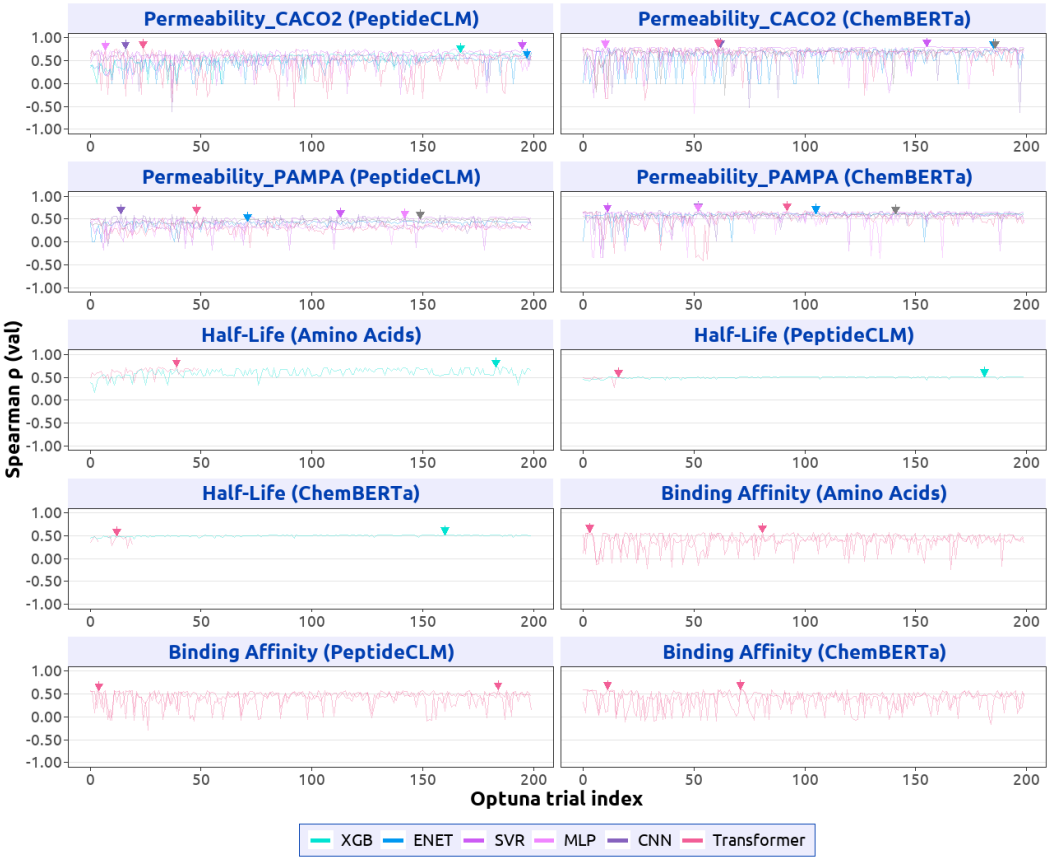

Figure S2: **Optuna optimization traces for regression models.** The arrow marks the selected trial with the best Spearman correlation (ρ) for each run. Machine-learning-based algorithms (XGB, ENET, SVR) were optimized for 200 trials, and neural-network-based algorithms (MLP, CNN, Transformers, where applicable) were optimized for 50 trials. SMILES tokenization results in longer input sequences and, therefore, slower training for neural-network-based models. The model choices for half-life and binding affinity reflect computational constraints and design considerations.

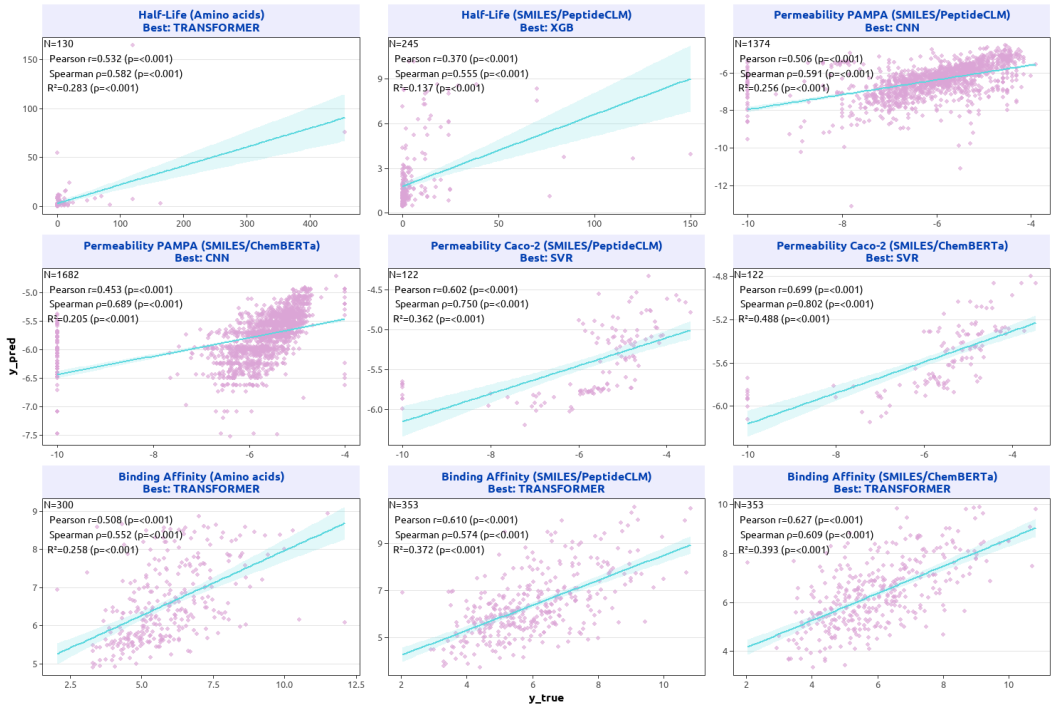

Figure S3: **Top model performance for regression models.** The best-performing models were selected and refit on the validation set for final evaluation. For half-life prediction, models were trained using cross-validation. The correlations are reported based on out-of-fold predictions collected during training. Specifically, in k-fold cross-validation, each fold's predictions were generated by models trained on the remaining $k-1$ folds, and the final results aggregate all held-out predictions. Reported statistics are computed on a single run with random seed 1986; mean values reported in Table S3 are averaged across bootstrap resamples and may differ slightly.

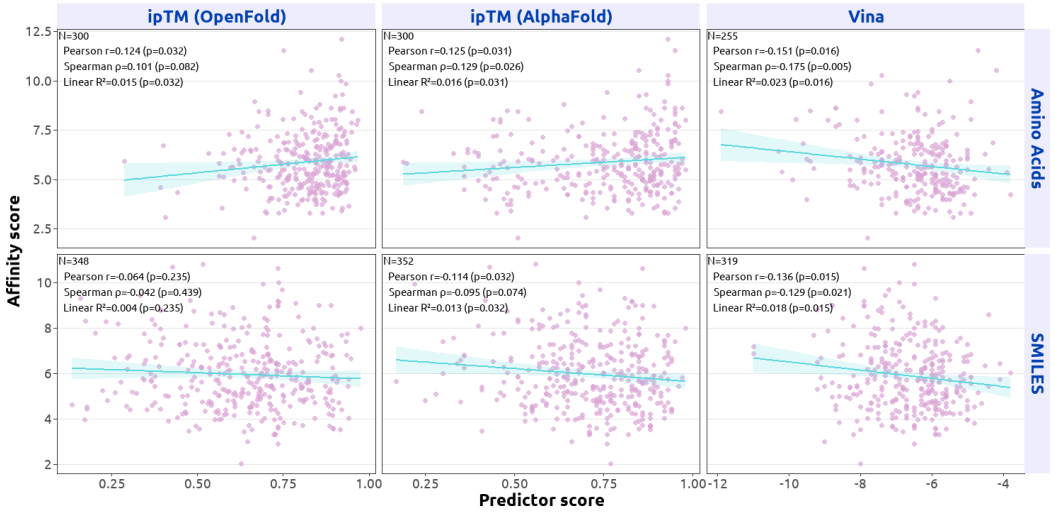

Figure S4: **Correlation between normalized experimental affinity score, OpenFold3 ipTM score, AlphaFold3 ipTM score, and the VINA docking score.**

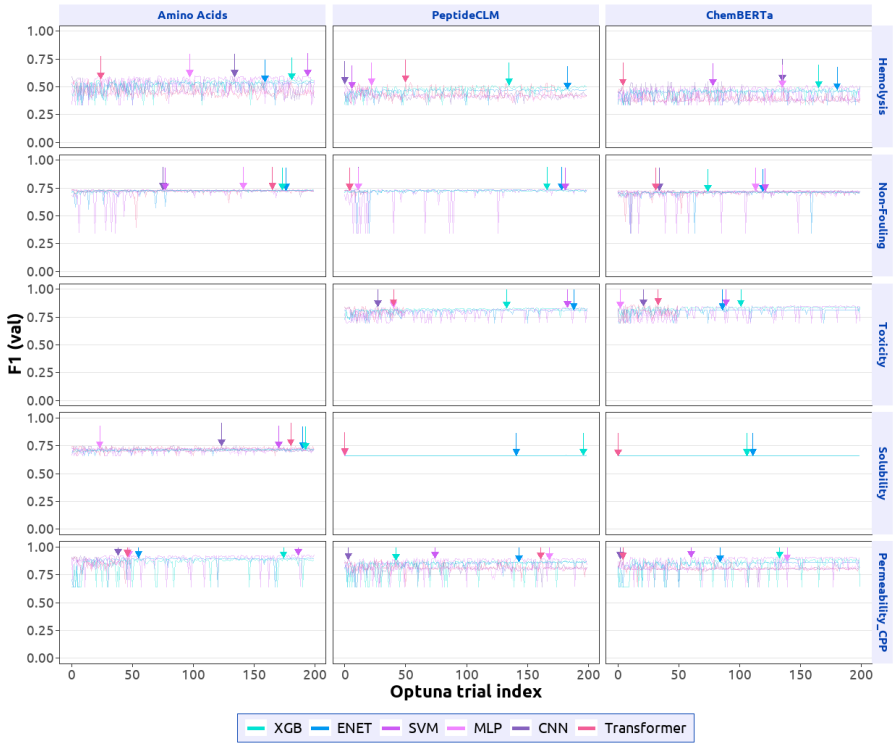

Figure S5: **Optuna optimization traces.** The arrow marks the selected trial with best F1 value for each run. All models were optimized for 200 trials.

Table S2: **95% Confidence Intervals for Classification Metrics.** Best F1 scores across all embeddings for each property are **bolded**. ML models (XGBoost, SVM, ElasticNet): mean [95% bootstrap percentile CI, n=2000]; time is single refit (s). DNN models (MLP, CNN, Transformer): mean ± std over 5 seeds; time in seconds. Best model on F1 per property per embedding type is bolded. The result matches Figure 3. For solubility, all models perform similarly.

Property	XGBoost			SVM			ElasticNet		
	F1	AUC	Time	F1	AUC	Time	F1	AUC	Time
ESM-2 embedding									
Hemolysis	0.55 [0.50, 0.60]	0.80 [0.77, 0.83]	3	**0.60 [0.55, 0.64]**	0.82 [0.78, 0.84]	1	0.54 [0.49, 0.59]	0.77 [0.74, 0.81]	1
Non-Fouling	0.73 [0.70, 0.75]	0.94 [0.93, 0.95]	9	0.73 [0.71, 0.76]	0.94 [0.93, 0.95]	2	0.73 [0.70, 0.75]	0.94 [0.93, 0.94]	2
Solubility	0.73 [0.71, 0.74]	0.78 [0.76, 0.79]	3	0.73 [0.72, 0.75]	0.79 [0.77, 0.80]	4	0.72 [0.71, 0.74]	0.78 [0.76, 0.79]	1
$\text{Permeability}_{\text{CPP}}$	0.89 [0.86, 0.92]	0.96 [0.94, 0.97]	3	**0.93 [0.90, 0.95]**	0.96 [0.94, 0.98]	7	0.91 [0.88, 0.93]	0.96 [0.94, 0.97]	1
PeptideCLM embedding									
Hemolysis	0.49 [0.44, 0.53]	0.75 [0.72, 0.78]	2	0.49 [0.44, 0.53]	0.75 [0.71, 0.78]	2	0.47 [0.43, 0.52]	0.71 [0.68, 0.75]	1
Non-Fouling	0.72 [0.70, 0.75]	0.93 [0.93, 0.94]	2	0.73 [0.71, 0.76]	0.94 [0.93, 0.94]	1	**0.73 [0.71, 0.76]**	0.93 [0.93, 0.94]	1
$\text{Permeability}_{\text{CPP}}$	0.88 [0.85, 0.91]	0.95 [0.93, 0.96]	1	0.90 [0.87, 0.93]	0.94 [0.92, 0.96]	1	0.87 [0.83, 0.90]	0.93 [0.90, 0.95]	2
Toxicity	0.83 [0.81, 0.84]	0.90 [0.89, 0.91]	3	0.84 [0.82, 0.85]	0.89 [0.88, 0.91]	1	0.81 [0.80, 0.83]	0.87 [0.86, 0.89]	1
Solubility	0.66 [0.65, 0.68]	0.57 [0.55, 0.59]	46	0.66 [0.66, 0.66]	0.60 [0.60, 0.60]	146	0.66 [0.66, 0.66]	0.61 [0.61, 0.61]	1090
ChemBERTa embedding									
Hemolysis	0.47 [0.42, 0.52]	0.74 [0.71, 0.78]	1	0.51 [0.46, 0.56]	0.76 [0.72, 0.79]	1	0.46 [0.42, 0.51]	0.71 [0.67, 0.75]	3
Non-Fouling	0.71 [0.69, 0.74]	0.93 [0.92, 0.94]	1	0.72 [0.70, 0.75]	0.93 [0.92, 0.94]	36	0.72 [0.69, 0.74]	0.93 [0.92, 0.94]	1
$\text{Permeability}_{\text{CPP}}$	0.89 [0.86, 0.92]	0.95 [0.93, 0.97]	4	**0.91 [0.88, 0.94]**	0.96 [0.95, 0.98]	1	0.86 [0.83, 0.90]	0.93 [0.90, 0.95]	3
Toxicity	0.83 [0.81, 0.84]	0.90 [0.89, 0.91]	3	**0.85 [0.84, 0.87]**	0.91 [0.89, 0.92]	1	0.81 [0.80, 0.83]	0.88 [0.86, 0.89]	2
Solubility	0.66 [0.65, 0.68]	0.54 [0.52, 0.56]	5	0.66 [0.66, 0.66]	0.55 [0.55, 0.55]	22	0.66 [0.66, 0.66]	0.55 [0.55, 0.55]	116

Property	MLP			CNN			Transformer		
	ρ	RMSE	Time (s)	ρ	RMSE	Time (s)			
ESM-2 embedding									
Hemolysis	$0.5734_{\pm 0.0035}$	$0.8133_{\pm 0.0059}$	$4_{\pm 0}$	$0.5700_{\pm 0.0096}$	$0.8090_{\pm 0.0081}$	$4_{\pm 1}$	$0.5722_{\pm 0.0067}$	$0.8174_{\pm 0.0054}$	$4_{\pm 1}$
Non-Fouling	$0.7341_{\pm 0.0015}$	$0.9374_{\pm 0.0006}$	$2518_{\pm 469}$	$0.7324_{\pm 0.0026}$	$0.9365_{\pm 0.0040}$	$2267_{\pm 820}$	$\mathbf{0.7361_{\pm 0.0039}}$	$0.9378_{\pm 0.0012}$	$2119_{\pm 127}$
Solubility	$0.7303_{\pm 0.0021}$	$0.7902_{\pm 0.0015}$	$4_{\pm 1}$	$\mathbf{0.7491_{\pm 0.0031}}$	$0.8183_{\pm 0.0026}$	$4_{\pm 0}$	$0.7469_{\pm 0.0023}$	$0.8170_{\pm 0.0039}$	$4_{\pm 0}$
$\text{Permeability}_{\text{CPP}}$	$0.9071_{\pm 0.0051}$	$0.9583_{\pm 0.0009}$	$5_{\pm 2}$	$0.9136_{\pm 0.0048}$	$0.9609_{\pm 0.0037}$	$7_{\pm 3}$	$0.9071_{\pm 0.0091}$	$0.9606_{\pm 0.0045}$	$4_{\pm 0}$
PeptideCLM embedding									
Hemolysis	$0.5000_{\pm 0.0035}$	$0.7535_{\pm 0.0051}$	$4_{\pm 0}$	$0.5061_{\pm 0.0116}$	$0.7625_{\pm 0.0121}$	$4_{\pm 0}$	$0.5120_{\pm 0.0146}$	$0.7652_{\pm 0.0119}$	$4_{\pm 0}$
Non-Fouling	$0.7316_{\pm 0.0020}$	$0.9362_{\pm 0.0004}$	$8181_{\pm 1072}$	$0.7334_{\pm 0.0018}$	$0.9366_{\pm 0.0004}$	$8766_{\pm 2277}$	$0.7289_{\pm 0.0020}$	$0.9358_{\pm 0.0002}$	$7820_{\pm 956}$
$\text{Permeability}_{\text{CPP}}$	$0.8863_{\pm 0.0078}$	$0.9429_{\pm 0.0079}$	$897_{\pm 259}$	$0.8815_{\pm 0.0036}$	$0.9405_{\pm 0.0027}$	$962_{\pm 133}$	$0.8779_{\pm 0.0067}$	$0.9345_{\pm 0.0072}$	$790_{\pm 265}$
Toxicity	$0.8342_{\pm 0.0023}$	$0.8958_{\pm 0.0034}$	$3597_{\pm 590}$	$0.8443_{\pm 0.0033}$	$0.9056_{\pm 0.0016}$	$3398_{\pm 721}$	$0.8469_{\pm 0.0043}$	$0.9069_{\pm 0.0052}$	$2705_{\pm 248}$
Solubility	$0.6633_{\pm 0.0004}$	$0.5487_{\pm 0.0109}$	$26598_{\pm 8939}$	$0.6661_{\pm 0.0017}$	$0.5994_{\pm 0.0254}$	$48924_{\pm 16132}$	$\mathbf{0.6680_{\pm 0.0010}}$	$0.6321_{\pm 0.0057}$	$31500_{\pm 7756}$
ChemBERTa embedding									
Hemolysis	$0.4861_{\pm 0.0094}$	$0.7287_{\pm 0.0109}$	$1372_{\pm 446}$	$\mathbf{0.5295_{\pm 0.0094}}$	$0.7817_{\pm 0.0083}$	$2054_{\pm 436}$	$0.4922_{\pm 0.0190}$	$0.7356_{\pm 0.0234}$	$1487_{\pm 507}$
Non-Fouling	$0.7263_{\pm 0.0029}$	$0.9337_{\pm 0.0005}$	$2861_{\pm 466}$	$0.7265_{\pm 0.0009}$	$0.9352_{\pm 0.0005}$	$3579_{\pm 698}$	$0.7267_{\pm 0.0013}$	$0.9351_{\pm 0.0007}$	$3191_{\pm 1039}$
$\text{Permeability}_{\text{CPP}}$	$0.8527_{\pm 0.0308}$	$0.9156_{\pm 0.0276}$	$759_{\pm 571}$	$0.8959_{\pm 0.0050}$	$0.9468_{\pm 0.0049}$	$1345_{\pm 196}$	$0.8906_{\pm 0.0072}$	$0.9426_{\pm 0.0090}$	$1060_{\pm 201}$
Toxicity	$0.8434_{\pm 0.0019}$	$0.9000_{\pm 0.0058}$	$5826_{\pm 405}$	$\mathbf{0.8529_{\pm 0.0044}}$	$0.9138_{\pm 0.0016}$	$5195_{\pm 967}$	$0.8527_{\pm 0.0020}$	$0.9113_{\pm 0.0037}$	$3178_{\pm 648}$
Solubility	$0.6623_{\pm 0.0001}$	$0.5208_{\pm 0.0008}$	$6491_{\pm 104}$	$0.6643_{\pm 0.0008}$	$0.5694_{\pm 0.0083}$	$18128_{\pm 6530}$	$0.6624_{\pm 0.0002}$	$0.5184_{\pm 0.0137}$	$6752_{\pm 846}$

Table S3: **95% Confidence Intervals for Regression Metrics Across Models and Embeddings.** Best Spearman rho (ρ) values across all embeddings for both properties are **bolded**. DNN models (MLP, CNN, Transformer): mean ± std over 5 random seeds. ML models (XGBoost, SVR, ElasticNet): mean [95% bootstrap percentile CI, n=2000]. Spearman ρ: bootstrap percentile CI. DNN training time reported as mean ± std over 5 seeds (seconds); ML time is a single refit.

Property	XGBoost			SVR			ElasticNet		
	ρ	RMSE	Time (s)	ρ	RMSE	Time (s)			
PAMPA (PeptideCLM)	0.51 [0.46, 0.55]	1.02 [0.95, 1.08]	4	0.54 [0.50, 0.58]	1.00 [0.93, 1.06]	12	0.45 [0.40, 0.50]	1.08 [1.02, 1.14]	61
PAMPA (ChemBERTa)	0.62 [0.58, 0.65]	1.02 [0.93, 1.10]	3	0.65 [0.61, 0.68]	1.04 [0.95, 1.13]	6	0.62 [0.59, 0.66]	1.01 [0.93, 1.09]	2
Caco-2 (PeptideCLM)	0.68 [0.54, 0.78]	1.24 [0.96, 1.53]	2	0.75 [0.65, 0.83]	1.30 [0.98, 1.62]	<1	0.56 [0.42, 0.68]	1.35 [1.04, 1.66]	2
Caco-2 (ChemBERTa)	0.77 [0.68, 0.84]	1.18 [0.89, 1.46]	1	**0.80 [0.72, 0.87]**	1.30 [1.01, 1.59]	<1	0.76 [0.68, 0.83]	1.28 [1.02, 1.55]	1
Property	**MLP**			**CNN**			**Transformer**		
	ρ	RMSE	Time (s)	ρ	RMSE	Time (s)			
PAMPA (PeptideCLM)	$0.4634_{\pm 0.0260}$	$1.1000_{\pm 0.0376}$	$5_{\pm 0}$	$0.5904_{\pm 0.0223}$	$1.0713_{\pm 0.0336}$	$5_{\pm 1}$	$0.5183_{\pm 0.0250}$	$1.3255_{\pm 0.0361}$	$6_{\pm 0}$
PAMPA (ChemBERTa)	$0.6533_{\pm 0.0158}$	$1.0210_{\pm 0.0453}$	$807_{\pm 142}$	$\mathbf{0.6883_{\pm 0.0146}}$	$0.9962_{\pm 0.0439}$	$1024_{\pm 250}$	$0.6701_{\pm 0.0149}$	$1.0156_{\pm 0.0433}$	$1191_{\pm 550}$
Caco-2 (PeptideCLM)	$0.7175_{\pm 0.0513}$	$1.4555_{\pm 0.1783}$	$6_{\pm 0}$	$0.7125_{\pm 0.0532}$	$1.3343_{\pm 0.1643}$	$6_{\pm 1}$	$0.7381_{\pm 0.0461}$	$1.2180_{\pm 0.1375}$	$6_{\pm 1}$
Caco-2 (ChemBERTa)	$0.7402_{\pm 0.0485}$	$1.2418_{\pm 0.1450}$	$172_{\pm 54}$	$0.7835_{\pm 0.0431}$	$1.3706_{\pm 0.1672}$	$94_{\pm 14}$	$0.7785_{\pm 0.0406}$	$1.2894_{\pm 0.1656}$	$96_{\pm 6}$

Table S4: **Final classification thresholds. Result reported at seed 1986.**

Properties	Best Model		Classification Thresholds		Best F1	
	Amino acid	SMILES	Amino acid	SMILES	Amino acid	SMILES
Hemolysis	SVM	CNN (ChemBERTa)	0.2521	0.564	0.596	0.543
Non-Fouling	Transformer	ENET (PeptideCLM)	0.57	0.6969	0.736	0.735
Solubility	CNN	Transformer (PeptideCLM)	0.377	0.329	0.749	0.668
Permeability_CPP	SVM	SVM (ChemBERTa)	0.5493	0.573	0.929	0.909
Toxicity	-	CNN (ChemBERTa)	-	0.49	-	0.86

Table S5: **95% Confidence Intervals for Half-Life Regression with different embeddings.** All results use bootstrap percentile CI (n=2000, CV-OOF).

Embedding	XGBoost		Transformer	
	ρ	RMSE	ρ	RMSE
Amino Acids (ESM-2-650M)	0.56 [0.41, 0.68]	37.41 [12.45, 59.66]	0.58 [0.44, 0.69]	39.77 [13.07, 63.57]
SMILES (PeptideCLM)	0.54 [0.43, 0.63]	14.68 [8.38, 20.51]	0.42 [0.30, 0.52]	16.01 [8.88, 22.66]
SMILES (ChemBERTa)	0.54 [0.43, 0.63]	14.13 [8.35, 19.57]	0.45 [0.32, 0.55]	17.24 [10.74, 23.57]
SMILES (ECFP)	0.51 [0.39, 0.61]	15.81 [8.59, 22.42]	0.47 [0.35, 0.58]	15.10 [8.30, 21.46]

Table S6: **Binding Affinity Regression Results** (mean ± std over 5 random seeds). Pooled: mean-pooled sequence embeddings. Unpooled: full token-level embeddings.

Embedding Strategy	ρ	RMSE	MAE	R^2
Pooled embeddings				
ESM-2 + ESM-2	$0.5576_{\pm 0.0091}$	$1.3767_{\pm 0.0634}$	$1.0109_{\pm 0.0533}$	$0.1695_{\pm 0.0761}$
ChemBERTa + ESM-2	$\mathbf{0.6056_{\pm 0.0127}}$	$1.3226_{\pm 0.0283}$	$1.0075_{\pm 0.0393}$	$0.2653_{\pm 0.0316}$
PeptideCLM + ESM-2	$0.5708_{\pm 0.0039}$	$1.3039_{\pm 0.0835}$	$1.0157_{\pm 0.0353}$	$0.2839_{\pm 0.0931}$
Unpooled embeddings				
ESM-2 + ESM-2	$0.5478_{\pm 0.0088}$	$1.4935_{\pm 0.2540}$	$1.1838_{\pm 0.2763}$	$0.0016_{\pm 0.3606}$
ChemBERTa + ESM-2	$\mathbf{0.5858_{\pm 0.0135}}$	$1.3517_{\pm 0.0477}$	$1.0460_{\pm 0.0303}$	$0.2322_{\pm 0.0538}$
PeptideCLM + ESM-2	$0.5758_{\pm 0.0172}$	$1.3626_{\pm 0.0396}$	$1.0516_{\pm 0.0206}$	$0.2200_{\pm 0.0452}$

Table S7: **Wild-type peptide classification with baseline embeddings.** Performance on wild-type classification tasks using baseline embeddings (one-hot, VHSE) vs. ESM-2-650M embeddings. VHSE Mean and One-Hot Mean embeddings are mean-pooled over sequence length. All models were trained for 200 Optuna trials.

Property	Model	Embedding	Best F1	Property	Model	Embedding	Best F1
Hemolysis	**SVM**	**ESM-2-650M**	**0.598**	Non-Fouling	**MLP**	**ESM-2-650M**	**0.739**
	CNN	One-Hot Unpooled	0.518		CNN	One-Hot Unpooled	0.735
	SVM	VHSE Mean	0.512		SVM	One-Hot Mean	0.723
	CNN	VHSE Unpooled	0.509		Transformer	One-Hot Unpooled	0.723
	SVM	One-Hot Mean	0.501		XGBoost	One-Hot Mean	0.721
	Transformer	VHSE Unpooled	0.501		Transformer	VHSE Unpooled	0.719
	XGBoost	One-Hot Mean	0.477		CNN	VHSE Unpooled	0.718
	Transformer	One-Hot Unpooled	0.466		MLP	One-Hot Unpooled	0.716
	XGBoost	VHSE Mean	0.449		SVM	VHSE Mean	0.708
	MLP	VHSE Unpooled	0.432		XGBoost	VHSE Mean	0.703
	Elastic Net	VHSE Mean	0.419		Elastic Net	One-Hot Mean	0.698
	MLP	One-Hot Unpooled	0.403		Elastic Net	VHSE Mean	0.695
	Elastic Net	One-Hot Mean	0.394		MLP	VHSE Unpooled	0.693
Property	Model	Embedding	Best F1	Property	Model	Embedding	Best F1
Permeability_CPP	**SVM**	**ESM-2-650M**	**0.929**	Solubility	**CNN**	**ESM-2-650M**	**0.754**
	CNN	One-Hot Unpooled	0.906		CNN	One-Hot Unpooled	0.682
	XGBoost	One-Hot Mean	0.896		CNN	VHSE Unpooled	0.679
	CNN	VHSE Unpooled	0.894		Transformer	VHSE Unpooled	0.672
	Transformer	One-Hot Unpooled	0.875		XGBoost	One-Hot Mean	0.670
	SVM	One-Hot Mean	0.871		SVM	VHSE Mean	0.667
	Transformer	VHSE Unpooled	0.869		Elastic Net	One-Hot Mean	0.666
	SVM	VHSE Mean	0.848		MLP	One-Hot Unpooled	0.666
	MLP	One-Hot Unpooled	0.836		MLP	VHSE Unpooled	0.666
	Elastic Net	One-Hot Mean	0.831		SVM	One-Hot Mean	0.666
	XGBoost	VHSE Mean	0.822		Transformer	One-Hot Unpooled	0.666
	MLP	VHSE Unpooled	0.813		XGBoost	VHSE Mean	0.666
	Elastic Net	VHSE Mean	0.794		Elastic Net	VHSE Mean	0.664

Table S8: **Binding affinity prediction with basic embeddings.** Performance on binding affinity regression using basic embeddings (one-hot, VHSE) vs. ESM-2-650M embeddings. VHSE Mean and One-Hot Mean embeddings are mean-pooled over sequence length. All models were trained for 200 Optuna trials.

Property	Model	Embedding	Spearman	Pearson	R^2
Binding Affinity	Transformer	**ESM-2-650M**	**0.557**	**0.512**	**0.310**
		One-Hot Mean	0.518	0.447	0.200
		One-Hot Unpooled	0.486	0.450	0.203
		VHSE Unpooled	0.471	0.200	0.040
		VHSE Mean	0.456	0.366	0.134

Table S9: **Half-life prediction with basic embeddings.** Performance on half-life regression using basic embeddings (one-hot, VHSE) vs. ESM-2-650M embeddings and ESMC embeddings. VHSE Mean and One-Hot Mean embeddings are mean-pooled over sequence length. All models were trained for 200 Optuna trials.

Property	Model	Embedding	Spearman	R^2
Half-Life	**XGBoost**	**ESMC**	**0.636**	**0.331**
	Transformer	ESMC	0.610	0.165
	Transformer	ESM-2-650M	0.582	0.310
	XGBoost	ESM-2-650M	0.558	0.252
	XGBoost	One-Hot Mean	0.533	0.148
	Transformer	VHSE Unpooled	0.383	0.070
	Transformer	One-Hot Unpooled	0.362	0.057
	XGBoost	VHSE Mean	0.371	0.031

Table S10: **Comparison of SMILES-based peptide embedding methods.** All models trained on ECFP embeddings are compared to the best-performing model for PeptideCLM and ChemBERTA embeddings. All models were trained for 200 Optuna trials.

Property	Model	Embedding	Best F1	Property	Model	Embedding	Best F1
Hemolysis	**CNN**	**ChemBERTa**	**0.5295**	Non-Fouling	**SVM**	**PeptideCLM**	**0.73**
	Transformer	PeptideCLM	0.5120		CNN	ECFP	0.7292
	SVM	ECFP	0.4791		Transformer	ChemBERTa	0.7267
	XGBoost	ECFP	0.4786		Transformer	ECFP	0.7260
	MLP	ECFP	0.4721		SVM	ECFP	0.7252
	Transformer	ECFP	0.4684		CNN	ECFP	0.7215
	CNN	ECFP	0.452		MLP	ECFP	0.7214
	Elastic NEt	ECFP	0.4281		SVM	ECFP	0.7184
Property	Model	Embeddings	Best ρ	Property	Model	Embeddings	Best ρ
Permeability_CACO2	**SVR**	**ChemBERTa**	**0.8**	Permeatiliby_PAMPA	**CNN**	**ChemBERTa**	**0.6883**
	XGBoost	ECFP	0.7987		MLP	ECFP	0.5977
	SVR	ECFP	0.7506		CNN	ECFP	0.5909
	CNN	PeptideCLM	0.75		CNN	PeptideCLM	0.5904
	MLP	ECFP	0.7369		XGBoost	ECFP	0.5902
	CNN	ECFP	0.7289		Transformer	ECFP	0.5726
	Elastic Net	ECFP	0.7221		SVR	ECFP	0.5222
	Transformer	ECFP	0.6808		Elastic Net	ECFP	0.5003
Property	Model	Embeddings	Best F1	Property	Model	Embedding	Best ρ
Toxicity	**SVM**	**ChemBERTa**	**0.85**	Binding Affinity	**Transformer**	**ChemBERTA**	**0.6056**
	Transformer	PeptideCLM	0.8469		Transformer	PeptideCLM	0.5758
	SVM	ECFP	0.8462		XGBoost	ECFP	0.5302
	XGBoost	ECFP	0.8460		MLP	ECFP	0.475
	Transformer	ECFP	0.8458		CNN	ECFP	0.4634
	CNN	ECFP	0.8453		SVR	ECFP	0.4471
	MLP	ECFP	0.8375		Transformer	ECFP	0.4395
	Elastic Net	ECFP	0.8139		Elastic Net	ECFP	0.3572

Table S11: **Wild-type peptide classification with ESM-C embeddings.** Embeddings were mean-pooled over sequence length for ML models and left unpooled for neural network models. All models were trained for 200 Optuna trials.

Hemolysis		**Non-fouling**		**Permeability CPP**		**Solubility**	
Model	F1	Model	F1	Model	F1	Model	F1
SVM	0.5866	Transformer	0.7368	CNN	0.9087	CNN	0.7650
XGBoost	0.5730	CNN	0.7355	Transformer	0.9043	Transformer	0.7621
CNN	0.5586	MLP	0.7346	SVM	0.9019	MLP	0.7423
Elastic Net	0.5428	SVM	0.7312	MLP	0.9016	XGBoost	0.7309
MLP	0.5347	Elastic Net	0.7303	XGBoost	0.8978	Elastic Net	0.7308
Transformer	0.5270	XGBoost	0.7300	Elastic Net	0.8806	SVM	0.7098

Table S12: **Comparison of full finetuning and LoRA for half-life prediction based on the pretrained stability model.** Transformer models were initialized from the stability predictor and adapted to the half-life task using either full finetuning or low-rank adaptation (LoRA). Performance is shown for predicting either raw half-life or log-transformed half-life.

Properties	Model	Finetuning Method	Target Type	Spearman	R^2
Half-Life	Transformer	**Full**	**log**	**0.582**	**0.283**
		Full	raw	0.376	0.021
		LoRA	log	0.522	0.292
		LoRA	raw	0.342	0.066

Table S13: **XGBoost hyperparameters for classification and regression tasks.** The ranges indicate Optuna search space boundaries. †Indicates log-uniform sampling in log space.

XGBoost Hyperparameters		
Hyperparameter	**Classification**	**Regression**
Objective	`binary:logistic`	`reg:squarederror`
Eval metric	`logloss`	`rmse`
Lambda (λ)	$[10^{-8}, 50]^{\dagger}$	$[10^{-10}, 10^{2}]^{\dagger}$
Alpha (α)	$[10^{-8}, 50]^{\dagger}$	$[10^{-10}, 10^{2}]^{\dagger}$
Gamma (γ)	$[0, 10]$	$[0, 10]$
Max depth	$[2, 15]$	$[2, 16]$
Min child weight	$[1, 500]$	$[10^{-3}, 500]^{\dagger}$
Subsample	$[0.5, 1.0]$	$[0.5, 1.0]$
Colsample by tree	$[0.3, 1.0]$	$[0.3, 1.0]$
Learning rate	$[10^{-3}, 0.3]^{\dagger}$	$[10^{-3}, 0.3]^{\dagger}$
Boosting rounds	$[50, 1500]$	$[50, 2000]$
Early stopping	$[20, 200]$	$[20, 200]$
Tree method	`hist`	`hist`

Table S14: **Elastic Net hyperparameters for classification and regression tasks.** †Indicates log-uniform sampling. The class weight and selection are categorical choices.

Elastic Net Hyperparameters		
Hyperparameter	**Classification**	**Regression**
Model type	`cuLogReg`, penalty="elasticnet"	`cuElasticNet`
Solver	`qn` (cuML)	cuML
C (inverse regularization)	$[10^{-4}, 10^{3}]^{\dagger}$	-
α (regularization)	-	$[10^{-8}, 10]^{\dagger}$
ℓ_1 ratio	$[0, 1]$	$[0, 1]$
Class weight	{`balanced`}	-
Selection	-	{`cyclic`, `random`}
Max iterations	$[200, 5000]$	$[1000, 20000]$
Tolerance	$[10^{-6}, 10^{-2}]^{\dagger}$	$[10^{-6}, 10^{-2}]^{\dagger}$

Table S15: **Support Vector Machine hyperparameters for classification and regression tasks.** †Indicates log-uniform sampling. γ only applies to non-linear kernels.

Support Vector Machine Hyperparameters		
Hyperparameter	**Classification**	**Regression**
Model type	`cuSVC (cuML)`	`SVR (scikit-learn)`
Kernel	{rbf, linear, poly, sigmoid}	{rbf, linear, poly, sigmoid}
C (regularization)	$[10^{-3}, 10^{3}]^{\dagger}$	$[10^{-3}, 10^{3}]^{\dagger}$
γ (kernel coefficient)	$[10^{-6}, 10]^{\dagger}$ or `scale`	$[10^{-6}, 10]^{\dagger}$ or `scale`
ϵ (tube width)	-	$[10^{-4}, 1]^{\dagger}$
Class weight	{`balanced`}	-
Max iterations	$[200, 5000]$	-
Tolerance	$[10^{-6}, 10^{-2}]^{\dagger}$	-

Table S16: **Neural network hyperparameters for classification and regression tasks.** All models share the same base hyperparameters, with additional hyperparameters introduced only where required by the specific architecture. †Indicats log-uniform sampling. ‡The Huber δ controls outlier sensitivity.

Neural Network Hyperparameters		
Hyperparameter	**Classification**	**Regression**
Shared parameters (all architectures)		
Learning rate	$[10^{-5}, 3 \times 10^{-3}]^{\dagger}$	$[10^{-5}, 3 \times 10^{-3}]^{\dagger}$
Weight decay	$[10^{-8}, 10^{-2}]^{\dagger}$	$[10^{-10}, 10^{-2}]^{\dagger}$
Dropout rate	$[0, 0.5]$	$[0, 0.5]$
Batch size	{16, 32, 64}	{16, 32, 64}
Loss function	Cross-entropy	{MSE, Huber}
Huber $\delta^{\ddagger}$	-	$[0.5, 5]^{\dagger}$
MLP-specific		
Hidden units	{256, 512, 1024, 2048}	{256, 512, 1024, 2048}
CNN-specific		
Channels	{128, 256, 512}	{128, 256, 512}
Kernel size	{3, 5, 7}	{3, 5, 7}
Number of layers	$[1, 4]$	$[1, 4]$
Transformer-specific		
Model dimension (d_{model})	{128, 256, 384}	{128, 256, 384}
Attention heads	{4, 8}	{4, 8}
Number of layers	$[1, 4]$	$[1, 4]$
FFN dimension	{256, 512, 1024, 1536}	{256, 512, 1024, 1536}

Table S17: **Binding affinity model hyperparameters.** A cross-attention architecture was used for peptide-protein binding prediction. †Indicates log-uniform sampling.

Binding Affinity Predictor Hyperparameters	
Hyperparameter	**Value / Range**
Optimizer settings	
Learning rate (AdamW)	$[10^{-5}, 3 \times 10^{-3}]^{\dagger}$
Weight decay	$[10^{-10}, 10^{-2}]^{\dagger}$
Architecture parameters	
Hidden dimension	{256, 384, 512, 768}
Attention heads	{4, 8}
Cross-attention layers	$[1, 4]$
Dropout rate	$[0, 0.4]$
Training configuration	
Batch size	{16, 32, 64, 128}
Auxiliary loss weight (λ_{cls})	$[0.1, 2.0]^{\dagger}$
Primary loss	MSE (binding affinity)
Auxiliary loss	BCE (binding classification)
Max epochs	50
Early stopping patience	10