

SynNat-BERT: An interpretable dual-modal BERT framework for blood–brain barrier permeability prediction and novel scaffold discovery

Junlin Dong

Shenzhen Institute for Drug Control

Shaoxin Huang

Shenzhen Institute for Drug Control

Suyi Liu

Shenzhen Institute for Drug Control

Siqing Chen

Shenzhen Institute for Drug Control

Zhen Zhang

Shenzhen Institute for Drug Control

Dao Zeng

Shenzhen Institute for Drug Control

Yuan Ji

Shenzhen Institute for Drug Control

Shuguang Yuan

Chinese Academy of Sciences

Horst Vogel

Chinese Academy of Sciences

Xie-an Yu

Shenzhen Institute for Drug Control

Ying Tan

`tan.ying@sz.tsinghua.edu.cn`

Tsinghua University

Bing Wang

Shenzhen Institute for Drug Control

Research Article

Keywords: Blood-brain barrier permeability, interpretable dual-modal BERT framework, substructural tokenization, molecular contextualization, scaffold discovery

Posted Date: September 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10742289/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

SynNat-BERT: An interpretable dual-modal BERT framework for blood-brain barrier permeability prediction and novel scaffold discovery

Junlin Dong^{1,2,†}, Shaoxin Huang^{3,†}, Suyi Liu^{1,†}, Siqing Chen¹, Zhen Zhang¹, Dao Zeng¹, Yuan Ji¹,
Shuguang Yuan^{3,4}, Horst Vogel^{3,5}, Xie-an Yu^{1,*}, Ying Tan^{2,*}, Bing Wang^{1,*}

¹ Shenzhen Key Laboratory of Drug Quality Standard Research, Shenzhen Institute for Drug Control,
Shenzhen, 518057, Guangdong, China

² Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, Guangdong,
China

³ Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, 518055,
Guangdong, China

⁴ AlphaMol Science Ltd, Shenzhen, 518055, Guangdong, China

⁵ Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne, Lausanne,
Switzerland

*To whom all correspondence should be addressed: yuxieanalj@126.com (Xie-an Yu);
tan.ying@sz.tsinghua.edu.cn (Ying Tan); wangbingszyj@163.com (Bing Wang)

† These authors contributed equally to this work.

Abstract

Blood-brain barrier permeability (BBBP) prediction is hindered by drug-centric assumptions and limited chemical diversity, leading to poor generalization and unreliable quantitative ranking. We present SynNat-BERT, an interpretable dual-modal self-supervised framework that learns transferable molecular representations from 1.7 million unlabeled molecules spanning synthetic and natural-product chemical space - two orders of magnitude fewer than many over 100-million-scale molecular language or multimodal models. SynNat-BERT combines a substructure-aware tokenizer with transformer contextualization and adapts to BBBP by integrating Gaussian-reweighted, interpretable physicochemical descriptors, avoiding opaque, high-dimensional fingerprint batches. This design supports accurate continuous BBBP scoring and calibrated ranking, while enabling quantitative attribution of property-level contributions. Under scaffold-based splits, SynNat-BERT improves both classification and regression (AUROC 0.891; Pearson r 0.677) over baselines and is more robust. Experiments and simulations validate alkaloid hits, with 3 of 5 candidates exhibited permeability comparable to, or even superior to, those of the positive controls, revealing previously unreported BBB-permeable scaffolds.

Scientific contribution

We present SynNat-BERT, an interpretable dual-modal self-supervised framework that learns transferable molecular representations from 1.7 million unlabeled molecules spanning synthetic and natural-product chemical space. The framework combines a substructure-aware tokenizer with transformer contextualization and adapts to blood-brain barrier permeability dataset by integrating Gaussian-reweighted, interpretable physicochemical descriptors, avoiding opaque and high-dimensional fingerprint batches. Applying the quantitative SynNat-BERT for a proof-of-concept virtual screening and ranking, we identify novel permeable candidates comparable to, or even superior to positive control as validated by Transwell assay.

Keywords

Blood-brain barrier permeability, interpretable dual-modal BERT framework, substructural tokenization, molecular contextualization, scaffold discovery

Introduction

The blood-brain barrier (BBB) represents a critical challenge in central nervous system (CNS) drug design[1, 2]. Its highly selective physiological structure prevents approximately 98% of therapeutic agents from penetrating into the brain[3, 4, 5]. Experimental measurements of BBB permeability, as an early-stage data annotation strategy, is primarily conducted using either *in vitro* or *in vivo* approaches[6]. *in vivo* assessment of brain exposure or brain-to-plasma ratio (C_{brain}/C_{plasma}) is widely accepted as the most reliable reference for evaluating brain distribution[7]. In contrast, traditional approaches have long relied on high-throughput screening (HTS) via *in vitro* permeability assessments, including membrane-based permeability assays and cell-based transmembrane transport assays[8, 9]. However, such biological activity annotations are costly and time-consuming, rendering these approaches unsuitable for screening overly vast drug-like chemical space (estimated as around 10^{60})[10].

With continuous advances in computational power, employing artificial intelligence approaches to achieve large-scale molecular physiological property predictions has been feasible[11, 12]. Conventional machine learning approaches, such as support vector machines (SVM) and Bayesian sampling, model molecular fingerprints in bulk but often suffer from limited feature relevance and poor generalization to unseen data[13, 14]. In contrast, deep learning models capture label-associated molecular representations in a data-driven, end-to-end manner, alleviating the reliance on sophisticated feature engineering[15, 16]. However, substantial knowledge gaps remain in existing methods predicting natural product molecular properties. Firstly, among the approximately 10^5 labeled molecular entries collected in MoleculeNet and B3DB, only about 1% are associated with experimentally determined log BB values[17, 18]. Secondly, perhaps most importantly, natural products possess abundant evolutionary information and occupy a unique chemical space, yet are often underexplored[19]. Consequently, existing deep learning models training on general chemical databases, such as ZINC and ChEMBL, demonstrate significant limitations towards unseen molecules across broader chemical space[20].

Recently, self-supervised learning (SSL) has emerged as a major paradigm in molecular property prediction, effectively addressing the problem of data scarcity, and capturing complex structural patterns of natural products[21, 22]. In the area of natural language processing (NLP), the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model has been extensively applied to encode discrete word embedding into continuous latent space[23]. In computational chemistry, the abundance of unlabeled simplified molecular-input line-entry system (SMILES) strings provides an ideal resource for pre-training, enabling unsupervised models to learn generalized molecular representations[24]. Through masked tokens or contrastive learning strategies, SSL frameworks acquire generalizable molecular representations by pre-training on large-scale unlabeled molecular databases and transfer these learned chemical knowledge to various downstream tasks[25, 26, 27].

Although these methods provide new insights into improving molecular property prediction accuracy under data scarcity scenarios, advances in this field exhibit certain shortcomings. First, existing approaches predominantly emphasize atom-level representations while overlooking molecular substructures and functional groups that exert pronounced effects on physicochemical properties[28, 29]. Atom- and character-based tokenizers fail to capture key chemically meaningful units, such as functional groups, thereby limiting the model’s structural comprehension and interpretability. Yet, previous studies have attempted to segment molecules into chemically meaningful fragments, these strategies typically rely on predefined rules, which constrain the model’s capacity to generalize across unseen regions of chemical space[30, 31]. Moreover, the accuracy of property prediction lies not only on the model’s perception of chemical structure but also on intrinsic physicochemical characteristics, such as hydrogen bonding, lipophilicity, molecular size and charge - key molecular descriptors that reflect overall chemical behavior[32]. Despite progress in structure-aware feature fusion, current methods remain limited in unifying deep molecular representations with interpretable physicochemical descriptors, in rationally selecting feature combinations for complex prediction tasks, and in delivering reliable quantitative predictions across diverse chemical space[33].

To tackle the aforementioned challenges, we propose Synthetic chemical and Natural product-Bidirectional Encoder Representation from Transformer (SynNat-BERT), a framework designed to construct foundational chemical language model and to achieve accurate prediction and interpretation on the blood-brain barrier permeability (BBBP) downstream task. SynNat-BERT employs a data-driven byte-pair encoding (BPE)-based SMILES tokenizer to represent chemically diverse molecules from 1 million PubChem synthetic

chemicals and 0.7 million COCONUT natural products to form a 1.7 million merged data set at the substructure level. Through comprehensive evaluations of molecular reconstruction accuracy and visualization of low-dimensional embeddings, BPE tokenizer demonstrate reliable performance over atomic tokenizer in recognizing molecular scaffolds. Furthermore, a dual-modal post-training adaptation framework is introduced to integrate 7 interpretable physicochemical molecular descriptors with contextual representations through a multi-head cross-attention fusion strategy. Ablation studies and feature engineering showed that the incorporated features were crucial to improve BBBP classification predictions, achieving AUROC and AUPRC, 0.891 and 0.961, respectively. In addition, SynNat-BERT demonstrates robust chemical feasibility in multiple interpretability analysis tasks, providing invaluable insights into (i) clustering natural products with similar scaffolds under low-dimensional manifold, (ii) enriching compounds satisfying BBB-permeable properties within the top ranking list, and (iii) elucidating structural determinants of BBB penetration via the attention mechanism. To further illustrate the capability of screening, we applied the regressive SynNat-BERT model to conduct a prospective virtual screening over the pre-training data set, where the model identified 4,590 potent BBB-penetrable alkaloid scaffolds distinct from the existing BBBP data sets. Finally, we test 5 alkaloid candidates with varying scaffolds for in-house validation via *in vitro* Transwell assay, leading to the discovery of 3 novel BBB-penetrable entities. Additionally, molecular dynamics simulations verifies the passive diffusion mechanism of these identified novel entities, further highlighting the flexible berbine scaffold as a key structural determinant of BBB penetration. In contrast, replacing the hydrogenated flexible scaffold with a rigid aromatic scaffold markedly reduced both the predicted BBB permeability scores and the experimentally measured permeability, in close agreement with the SynNat-BERT predicted attention maps. In summary, the proposed SynNat-BERT model achieved multifacet performances in BBBP property prediction and interpretation. The accompanying mechanistic analyses provide theoretical insights and practical utilities for artificial intelligence-driven molecular property prediction and lead optimization.

Results and discussion

Overview of SynNat-BERT framework

As illustrated in Fig. 1, SynNat-BERT is an encoder-only transformer model designed to predict and interpret real-world blood-brain barrier permeability. Based on the RoBERTa (Robustly optimized Bidirectional Encoder Representation from Transformer approach) transformer implementation[34], SynNat-BERT takes unlabeled small molecule strings as inputs and outputs the blood-brain barrier permeability score as well as the substructure-level attention weights.

To address the issues arising from data scarcity and tokenization biases, SynNat-BERT leverages a self-supervised architecture to encode chemically diverse molecules with a BPE substructural tokenizer. The base encoder, modeled with stacked self-attention layers, constructs a chemical foundation model via stochastically masking BPE tokenizer, whereas the post-training adaption employs a multi-head cross-attention fusion mechanism to achieve dual-modality alignment between molecular context and physiochemical fingerprints. Each molecular context is represented as a ‘sentence’ constructed by various substructure tokens, along with a positional encoding layer to generate a mean pooling embedding vector. Each molecular descriptor depicting physiochemical fingerprint is transformed by quantile ranking and linearly projected into the same representational space of textual embedding (Methods).

To validate the generalizability and chemical interpretability of the model, we performed a sequential perspective and retrospective analysis on pre-training distribution, unseen screening, attention visualization, and experimental correlation, to benchmark the performance of SynNat-BERT against other models. Besides, we applied structural novelty post hoc analysis and feature engineering strategy to validate the multi-head cross-attention fusion model as a robust tool demonstrating real-world utilities in expanding BBB-permeable scaffolds and identifying BBB-penetrable candidates.

Performance of SynNat-BERT pre-training natural product-enhanced BPE tokenizer

Firstly, to evaluate the representational capacity of the pre-trained BPE tokenizer for natural products, we compared the pre-training performance using token recovery accuracy as the evaluation metrics (Methods).

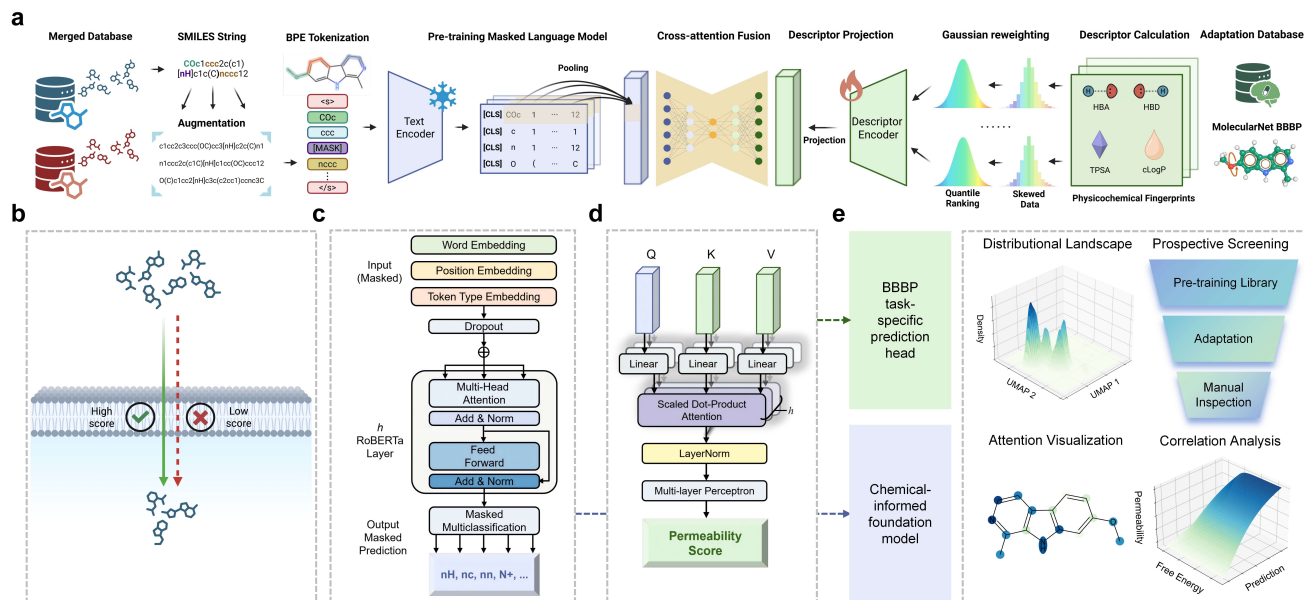


Figure 1: **Overview of SynNat-BERT framework.** (a) Pre-training and post-training pipelines of SynNat-BERT. Text encoder (left panel) is integrated with descriptor encoder (right panel) via a cross-attention fusion mechanism. (b) Schematic diagram of SynNat-BERT downstream validation. The output scores of the adapted SynNat-BERT are used to calibrated with molecular dynamics calculated free energy surface and Transwell measured permeability. (c) The architecture of SynNat-BERT self-attention text encoder. Masked language modeling strategy is leveraged to pre-train the model by outputting the most probable tokens based on bidirectional SMILES context. (d) The framework of cross-attention fusion network. The dual-modal input is encoded by text query (Q), descriptor key (K), and descriptor value (V). Multi-head cross-attention mechanism is used ($h=6$). (e) Evaluation protocols used to benchmark SynNat-BERT’s performance.

SMILES Tokenizer BERT (ST-BERT) adopted the same settings and was pre-trained based on the widely used atom-level tokenization[35]. Herein, SMILES enumeration strategy was applied to augment pre-training data set which potentially increase molecular diversity (Supplementary Fig. 1). As shown in Fig. 2a (left panel), BPE-BERT pre-trained on 1.7 million merged data set demonstrated the highest recovery rate, showing robustness over ST-BERT and models pre-trained on other scales of merged data set (Supplementary Tab. 1). In addition, to demonstrate the validity of integrating two data sets, we visualized the distributional differences between compounds from chemical databases, such as ChEMBL (2.8 M), ZINC (0.25 M) and PubChem (1M), and molecules derived from natural product database, i.e. COCONUT (0.7M, see Methods). The low-dimensional embeddings revealed that natural products occupied a more distinct chemical space, focusing on the outside regions of chemically synthetic space, suggesting that they could serve as a complement to synthetic compound libraries (Fig. 2b). Overall, these results indicated that the diversity of natural product compounds supplements the structural bias of chemically synthesized compounds, supporting our further selection of the 1.7 million pre-training data set.

Subsequently, we evaluated recovery accuracy differences of pre-training 1.7 M PubChem+COCONUT data sets using three strategies: (1) warm-start: internal validations of pre-training data sets to examine capability to recover pre-training molecules; (2) cold-data: external validations of pre-training data sets to test generalizability to recover unseen molecules from different sources; (3) cold-train: external validations using ChEMBL data set comparing other RoBERTa frameworks without pre-training. Our benchmark focused on the impact of integrating natural product-based molecules to the overall pre-training performances. Thus, we compared the integration of COCONUT with other chemical databases. Under warm-start setting, models pre-training with COCONUT data set consistently outperformed that without integrating COCONUT data set (PubChem: 0.803 $\rightarrow$ 0.845, ZINC: 0.409 $\rightarrow$ 0.497, and ChEMBL: 0.012 $\rightarrow$ 0.027). Although model pre-trained on the ChEMBL data set demonstrated slightly superior performance under the cold-data scenario, it may be attribute to the larger scale of ChEMBL data set, and its overlap with other data sets could potentially result in test data leakage. In addition, pre-training PubChem+COCONUT 1.7M data set consistently outperformed state-of-the-art (SOTA) RoBERTa baselines, particularly in the challenging external validation task (Fig. 2c). Our pre-trained model outperformed open-source models pre-trained on 10, 77, and 100 million molecules from the PubChem database in terms of reconstruction accuracy, i.e., PubChem10M_SMILES_BPE_450k ($P = 0.0016$), ChemBERTa-77M-MTR ($P = 1.27 \times 10^{-6}$), and ChemBERTa-100M-MLM ($P = 0.0021$), respectively (Supplementary Tab. 2). In addition, we observed molecules sharing similar substructures tended to cluster into specific regions, compared to unpretrained ChemBERTa-100M-MLM model (Fig. 2f) and SynNat-BERT models pre-trained using other hyperparameter settings (Supplementary Fig. 2). We attributed this success to the usage of masked BPE tokenizer with a diverse set of pre-training data set, which effectively captured substructure patterns in an unsupervised manner[36].

Incorporating interpretable descriptors enhanced BBBP predictions

Understanding drug-centric mechanistic knowledge is fundamental to investigate permeable scaffolds. Recent advances in descriptor-enhanced importance analysis enable the detailed profiling for molecular properties concerning brain permeation, thus, supporting the accurate establishment of CNS-targeted quantitative structure-activity relationship (QSAR) model[37, 38]. The complexity extracting bulk molecular features, including atom-bond descriptors, 2D/3D representations, and Morgan fingerprints, presents significant challenges developing CNS-targeted multi-parameter optimization (MPO) approaches[39, 40]. In this study, we incorporated 7 interpretable physicochemical descriptors, well established in previous report, within a cross-attention multi-modal fusion strategy (Supplementary Note. 1). The incorporated interpretable features were categorized into six sub-classes, including: (1) hydrogen bonding (NumHDonors, and NumHAcceptors)[41]; (2) lipophilicity (MolLogP)[42]; (3) molecular size and volume (MolWt)[43]; (4) molecular heteroatoms (NOCount)[44, 45]; (5) molecular shape and flexibility (NumRotatableBonds)[46]; (6) molecular topology (TPSA)[47].

Firstly, we performed ablation studies to evaluate the reliability of incorporating these descriptors. As illustrated in Fig. 3a, removing any of the 7 descriptors would result in declines in both AUROC and AUPRC, thereby impairing the overall classification performance. To validate the reliability of modality fusion strategy, we assessed the BBBP classification performances of model variants and visualized the

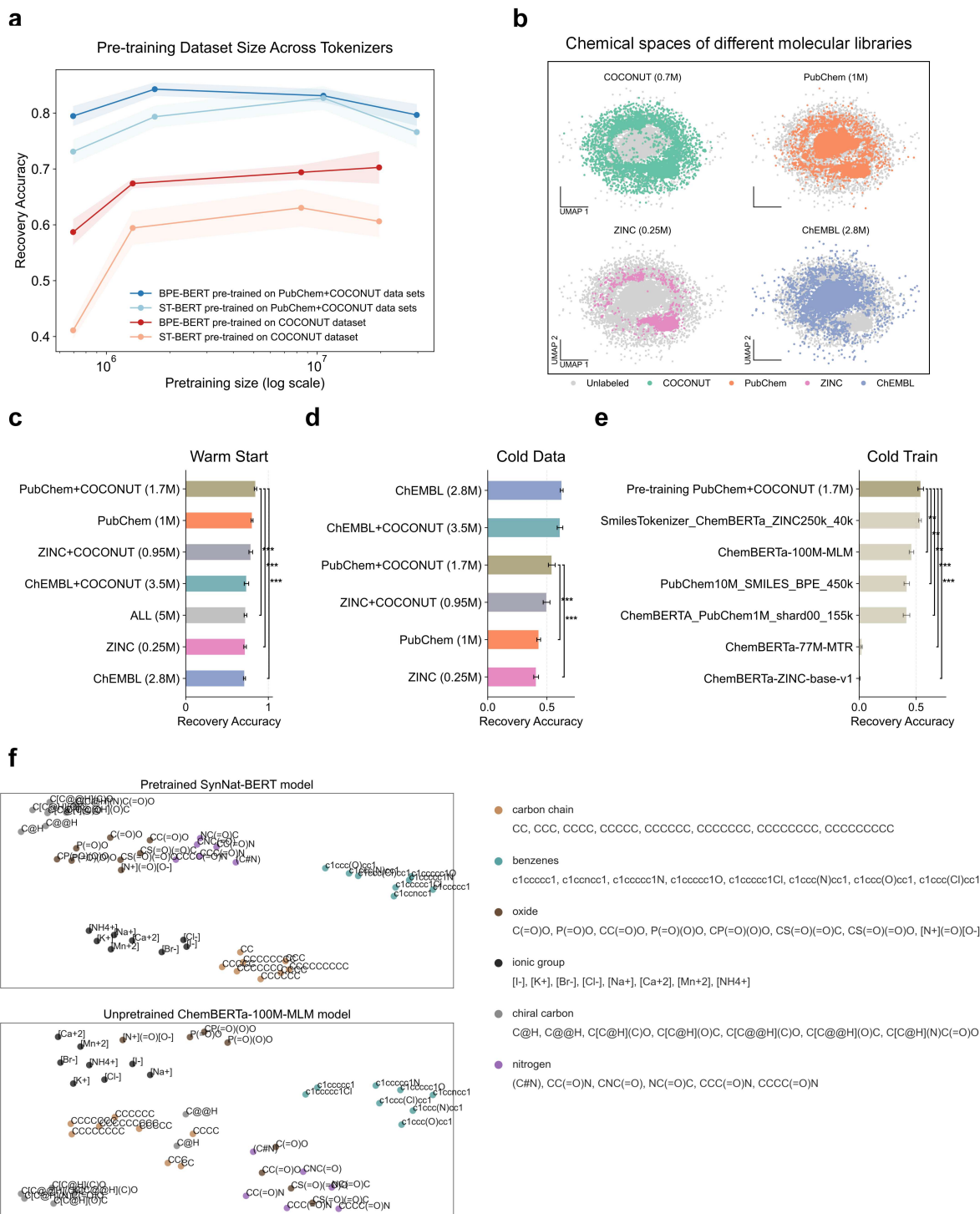


Figure 2: **Molecular generalizability of SynNat-BERT pre-training.** (a) The effect of pre-training data set size (left), and hyperparameters (right) as well as tokenizer selections to model recovery accuracy. (b) UMAP of molecular embeddings generated by SynNat-BERT, colored by different molecular libraries (green: COCONUT; orange: PubChem; purple: ZINC; blue: ChEMBL). (c-e) Recovery accuracy evaluations by different settings: warm-start (c), cold-data (d), and cold-train (e). For warm-start and cold-data scenarios, pre-training data set scales are labeled in corresponding brackets. Statistical significance is denoted as follows: $*P < 0.05$, $**P < 0.01$, $***P < 0.001$ (pairwise Welch’s t -test, two-tailed). (f) Distributions of substructure embeddings. Substructures are randomly chosen. The upper shows embeddings from 1.7 million compounds pre-trained SynNat-BERT, and the lower is generated by ChemBERTa-100M-MLM. Similar substructures are marked in the same colors.

attention weight distributions of each attention head (Supplementary Fig. 3 and Supplementary Tab. 3). In this study, SynNat-BERT adopted a cross-attention (CA) Gaussian-reweighted (GR) modality fusion framework with a frozen text encoder (FT). By ablating each model component, we observed that SynNat-BERT and SynNat-BERT (w/o FT) represented the most distinctive attention heads, exhibiting the lowest mean cosine similarity (0.026 ± 0.198 , Fig. 3b). The similarity distribution patterns between these two variants indicated that freezing the text encoder exerted minimal influence on dual-modal adaptation. In specific, attention head 2 (dark blue) preferentially focused on the hydroxyl group "O", representing hydrogen bond-related descriptors were highlighted by hidden neurons. Additionally, attention head 5 (light pink) concentrated on the contribution of unsaturated bonds, which implicitly encoded non-rotatable chemical bond information.

Furthermore, we comprehensively compared the physicochemical property distributions of the top-1000 molecules. Each model variant was adapted and applied to inference on the pre-training data set (1.7 M). Invalid, inorganic, and molecules with replicated scaffolds are filtered from the top-1000 ranking list. Beside 7 interpretable descriptors, we also compared quantitative estimate of druglikeness (QED), which measured the overall quality of the top-ranking molecules[49]. SynNat-BERT (w/o all) was excluded from visualization due to its significant variance from others. As a result, we observed that SynNat-BERT (w/o CA) had the most profound impact on the distributions of top-ranking molecules, leading to biased distribution patterns distinct from that of other model variants (Fig. 3c). Collectively, these results validated the chemical feasibility of incorporating interpretable descriptors through a Gaussian reweighted dual-modal fusion architecture, where the cross-attention mechanism effectively captured substructural patterns.

BBBP prediction performance compared with state-of-the-art MLs

To comprehensively evaluate our proposed SynNat-BERT, we selected six state-of-the-art ML models as baselines by inspecting BBBP prediction accuracy and low-dimensional embedding after finetuning. The first and second were conventional machine learning methods based on Morgan fingerprints, namely XGBoost and RandomForest, that consistently performed overwhelmingly in the MoleculeNet BBBP benchmarking data set. The third and fourth leveraged the widely used GNN models, in which the Attentive FingerPrint denoted as AttentiveFP[50] and the Directed Message-Passing Neural Network denoted as D-MPNN[51], respectively. In addition, two pretraining-adaptation architectures exploiting semantic information, namely SMILES-Pair Encoding BERT (SPE-BERT)[30] and Functional Group BERT (FG-BERT)[31] were implemented to verify the advances of BPE embedding and the cross-attention fusion adaptation strategy.

We took a typical scaffold of natural products, flavonoid, to illustrate different SMILES fragmentation led by varying tokenizers (Fig. 4a). In contrast to the widely applied SMILES-tokenizer, substructure-based tokenization paid more attention to involve chemical bond information. For instance, SPE facilitated the identification of "O=C" carbonyl bond, while Functional Group (FG) embedding split molecules into different substituents inevitably ignoring a number of fragments that have not been collated in the existing vocabulary. In this paper, we harnessed byte-pair encoding (BPE), a hybrid tokenization technique positioned between character- and word-level methods, to learn general and explainable representations of molecule structures. Such a tokenizer could not only capture contextualization related to the unsaturated aromatic bonds of benzene rings represented as a series of "c" in the lower case, e.g. "ccccc", but also exhibited distinct advantages in identifying branching patterns in polyheterocycles, such as a single "2" and "3" token, enabling it to distinguish between the digits and letters that denote ring-connection positions.

Fig. 4b depicted the ROC and PRC curves to assess the classification performance of SynNat-BERT compared to baseline models. SynNat-BERT outperformed all baseline models with AUROC and AUPRC scores of 0.891 ± 0.023 and 0.961 ± 0.008 , respectively (Supplementary Tab. 4). In addition, regression performances were also measured by training with the quantitative BBBP data set, where SynNat-BERT demonstrated outstanding performances in Pearson correlation coefficient (PCC), Spearman, and R^2 metrics with root mean squared error (RMSE) and mean absolute error (MAE) comparable to other baselines (Fig. 4c, Supplementary Fig. 4 and Supplementary Tab. 5). Although the model was primarily designed to classify BBB-permeable molecules, the performance of its regression variant further supported subsequent prospective virtual screening applications.

Moreover, we analyzed the physicochemical property distributions of the top-1000 molecules among all baselines. As revealed by previous studies, (i) MolWt < 450; (ii) TPSA < 90 ; (iii) 2 < MolLogP <

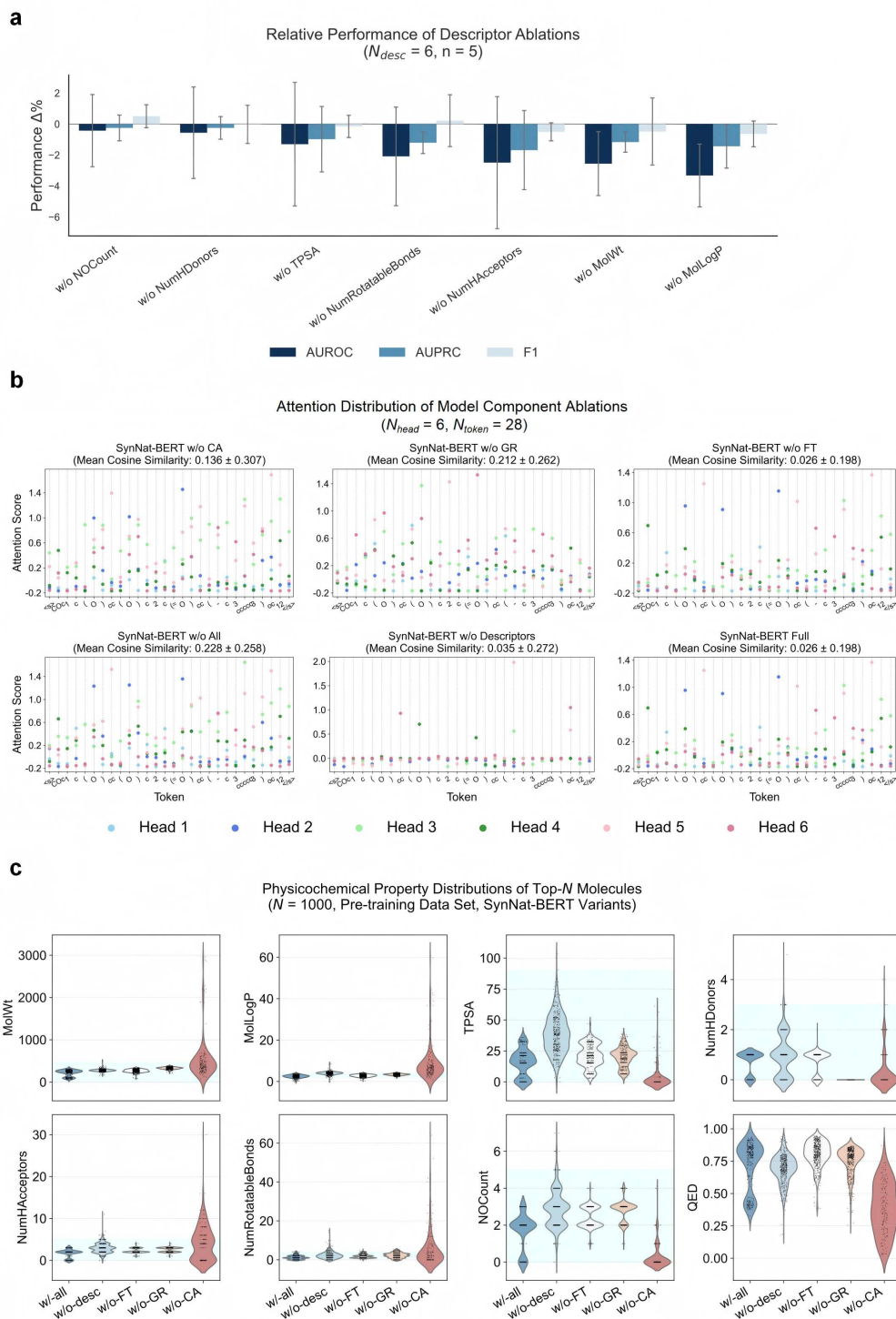


Figure 3: **Feature engineering of dual-modal fusion SynNat-BERT adapted to BBBP data set.** (a) Relative performance of SynNat-BERT by ablating each physicochemical descriptors. Errors derive from cross-validation adaptation ($n=5$). Relative performance is compared to the original SynNat-BERT model ($N_{desc} = 7$). (b) Attention-weight distributions of model variants, ablating each model component for an exemplified compound, wogonin, across six attention heads. Five model variants with the full SynNat-BERT model (SynNat-BERT full) are included: (1) w/o-desc: SynNat-BERT without descriptor modality; (2) w/o-FT: SynNat-BERT without freezing text; (3) w/o-GR: SynNat-BERT without Gaussian reweighted; (4) w/o-CA: SynNat-BERT without cross-attention fusion; (5) w/o-all: SynNat-BERT without the above components. Mean cosine similarity (m.c.s.) is quantified to compare the overall similarity between different attention heads (Methods). Errors derive from five-fold cross-validation adaptation ($n=5$). (c) Physicochemical properties of the top-1000 molecules screened by different model variants corresponding to panel b. Suggested ranges are colored by cyan reported by previous work[44, 48].

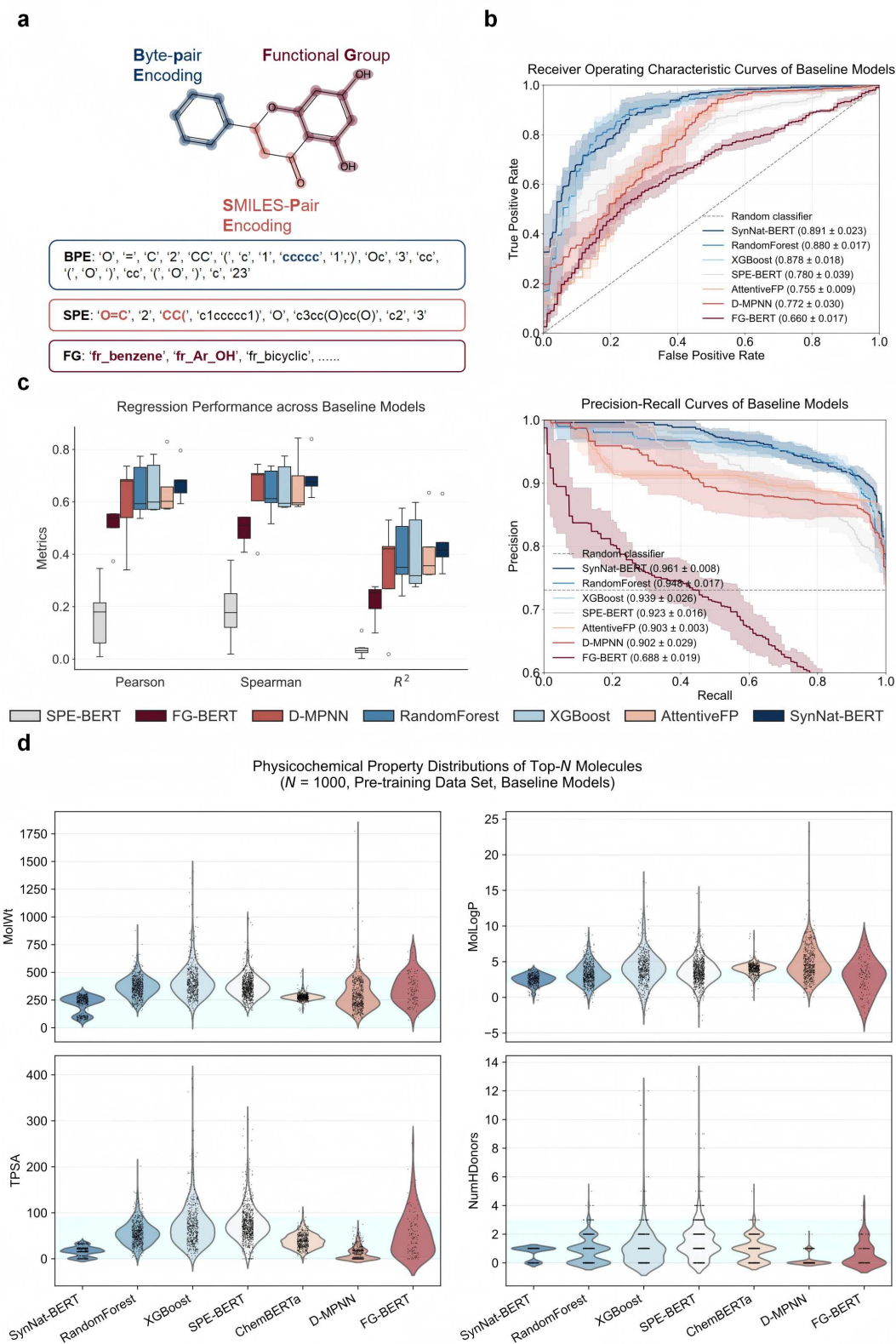


Figure 4: **Model comparison of SynNat-BERT with SOTA baselines.** (a) Schematic graph of different fragmentations in understanding molecular structures, colored by contextual tokenizers. (b) AUROC and AUPRC curves to assess the classification accuracy of different models. (c) Regression performance evaluated by three correlation metrics PCC, Spearman and R^2 across different baselines. Benchmarking models were trained on MoleculeNet BBBP regression data set ($n=1,058$). Colors of positive samples correspond to panel b. (d) Physicochemical properties of the top-1000 valid molecules screened by different baseline models. Suggested ranges are colored by cyan derived from [48]. Colors correspond to panel b.

4; and (iv) $\text{HBD} < 3$ were critical determinants of BBB permeability[48]. SynNat-BERT demonstrated robust advantages over other baselines, prioritizing molecules fulfilling the above suggested property ranges from the top ranking molecules (Fig. 4d and Supplementary Fig. 5). To summarize, SynNat-BERT not only demonstrated superior BBBP prediction accuracy but also excelled in screening capability over other competing models, demonstrating practical utility for drug-like molecular virtual screening.

A proof-of-concept virtual screening results in the identification of potent and distinct leads

After evaluating SynNat-BERT retrospectively on BBBP data set, we adopted the model in a prospective virtual screening settings (Fig. 5a). SynNat-BERT was trained on the real-world logBB quantitative data set (1,058 molecules), using a 95:5 train-validation split, with five cross-validation replicates to verify the robustness of the ensemble predictions (Methods). We applied the trained model to screen 1.7 million molecules from the pre-training library, a collection of synthetic compounds from PubChem and natural products from COCONUT databases (Methods). To expand the existing chemical space, we focused on exploring potential permeable hits derived from the library of natural products. In this study, we used classifiers from the COCONUT database, which provides access to classify natural products into 7 different categories, including terpenoids, alkaloids, shikimates and phenylpropanoids, fatty acids, polyketides, amino acids and peptides, and unknown (Methods). According to Supplementary Fig. 6, the regions occupied by BBB-permeable compounds highly overlapped with those of compounds receiving positive prediction scores, suggesting that molecules with prediction scores greater than 0 may possess BBB permeability[52]. Due to balanced performances across predicted logBB and maximum similarity metrics (second-best category among other natural products), we continued with 38,424 $\text{logBB} > 0$ alkaloids as underexplored and penetrable virtual hits (Fig. 5b and 5c).

For each virtual hit, we computed its synthetic accessibility score for brought in-house experimental validations. Among permeable hits, around 35% exhibited synthetic accessibility scores lower than 3, a threshold that ensured molecules can be obtained through simple reaction using commercial building blocks[53] (Fig. 5d). Given the broad range of Tanimoto similarities of alkaloid candidates, we initially selected candidates exhibiting structural dissimilarity to the training set ($\text{Tanimoto} < 0.4$), yielding 10,909 synthetically accessible and structurally dissimilar candidates[20, 54] (Fig. 5e, detailed threshold selection criteria see Supplementary Fig. 7). To alleviate filtering unfamiliar candidates within low-confidence regions, we considered testing candidates with similar scaffolds from CNS penetrable library which were used as external validations (SelleckChem, USA) (Fig. 5f). Among the 56 overlapped scaffolds, we manually selected 5 distinct clusters and filtered molecules with the highest SynNat-BERT scores (Methods). Since harmine was previously reported to transport across BBB monolayer cell, we applied it as a natural product-derived positive control[55]. The ability that the model prioritized potential candidates similar to the unseen CNS penetrable compounds demonstrated SynNat-BERT’s generalizability to BBB-penetrable substructures. Notably, as a proof-of-concept screening, additional permeable candidates that may exist in other categories, for instance, terpenoids and polyketides, sharing scaffold similarities with known CNS-penetrant compounds.

In vitro evaluation discovers BBB permeable scaffolds via passive diffusion mechanism

Given the filtered distinct clusters, a total of seven compounds – comprising five top-ranked candidates from the screening and two established positive controls (donepezil and harmine) – were selected for subsequent experimental evaluations. The experimentally measured apparent permeability was used to verify SynNat-BERT’s capability to prioritize permeable candidates and its predictivity of passive diffusion transmembrane mechanism that governs compound penetration.

Firstly, we applied an immortalized human brain endothelial cell line (hCMEC/D3) - astrocytes (AST) bilayer cell-cultured Transwell system to quantitatively measured the *in vitro* apparent permeability (P_{app}) of screened alkaloid candidates (Fig. 6a). The measured P_{app} of the positive control donepezil was $(3.36 \pm 1.07) \times 10^{-6}$ cm/s, which is around 5-time lower than a previously reported value $((15.6 \pm 3.1) \times 10^{-6}$ cm/s) measured within a monolayer system[56]. Another positive control, harmine, previously reported to cross the BBB, also showed a P_{app} of $(2.27 \pm 0.48) \times 10^{-6}$ cm/s above 10^{-6} cm/s in our bilayer co-cultured Transwell

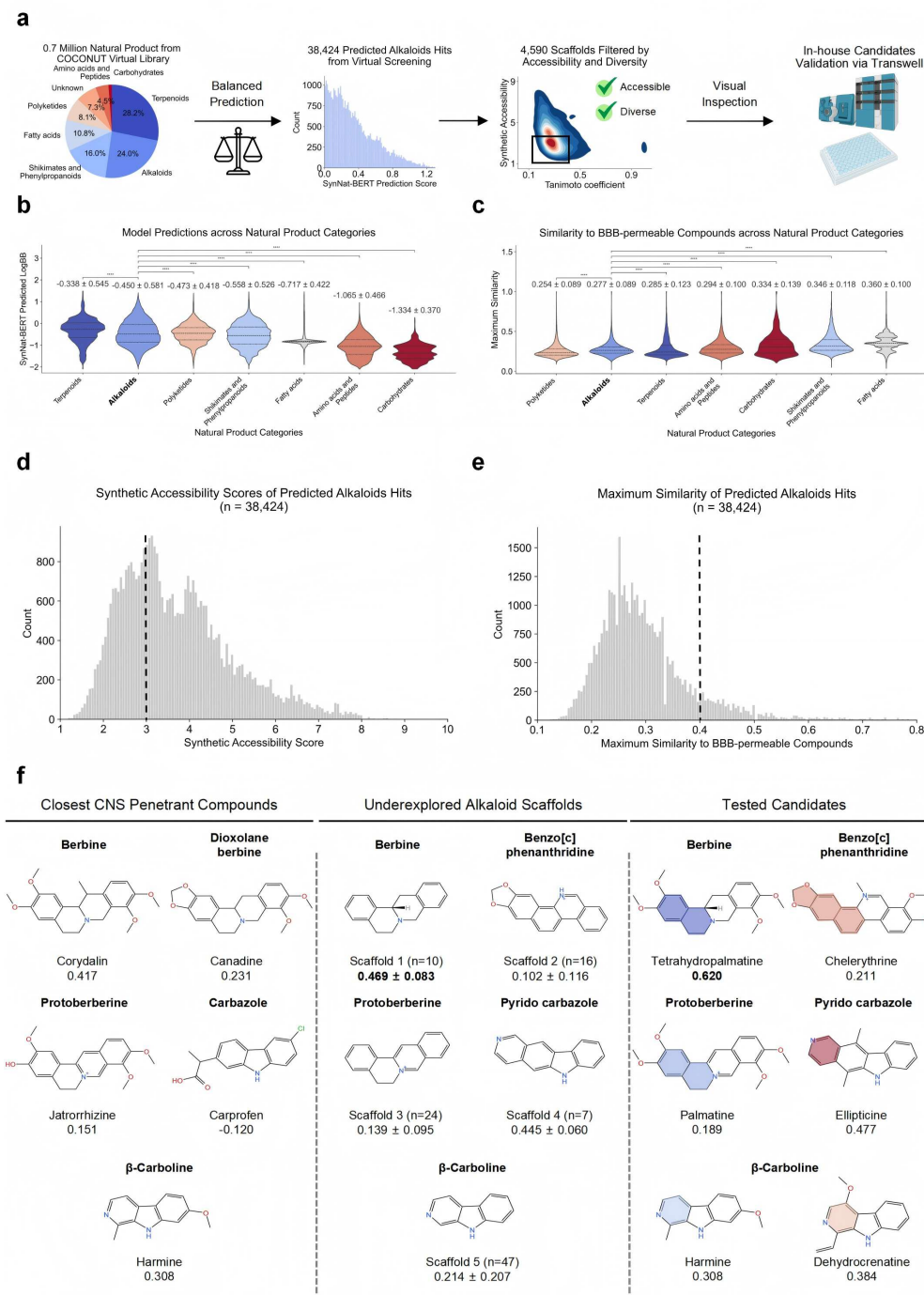


Figure 5: Prospective virtual screening of BBB-permeable natural products. (a) Virtual screening pipeline. SynNat-BERT model trained on the quantitative logBB data set is used to screen the entire pre-training space (1.7 M). Virtual hits are classified into varying natural product categories and are ranked by prediction scores (representing the permeability across BBB). Further filtering are based on multiple criteria, ensuring synthesizability and diversity from known BBB-permeable compounds. Finally, the selected library is brought in-house for hit confirmation via *in vitro* Transwell assay. This image was partially created with BioRender.com. (b-c) Violin plot of model predictions and maximum similarities across varying screened natural product categories. Mean value of each category is labeled. Statistical significance is denoted as follows: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, **** $P < 0.0001$ (pairwise Welch's *t*-test, two-tailed). (d-e) Distributions of synthetic accessibility scores and maximum similarities of alkaloids hits. Selection thresholds are indicated in dashed lines. (f) Visual inspection of the closest BBB-permeable compounds, underexplored alkaloid scaffolds and the tested candidates. SynNat-BERT scores are computed by the optimal screening model and labeled correspondingly. Key substitution groups are highlighted in the tested candidates with the basic scaffolds (isoquinoline, carbazole, and indole rings) colored by gray.

system[55, 57]. In addition, tetrahydropalmatine exhibited the highest permeability ($(3.58 \pm 0.48) \times 10^{-6}$ cm/s) among all tested hits, slightly exceeding that of the positive control donepezil, whereas ellipticine showed the lowest permeability (Fig. 6b). Other tested molecules, such as dehydrocrenatine and palmatine, exhibited P_{app} above 10^{-6} cm/s, indicating that 3 of 5 tested compounds were capable of penetrating the BBB (Supplementary Fig. 8 and Supplementary Tab. 6).

To further investigate the penetration mechanism of tested candidates and substructure details that govern the penetration process, we explored intermolecular interactions between the tested molecules and lipid transmembrane bilayers in a three-dimensional dynamic environment using molecular dynamics simulations[52]. SynNat-BERT prediction scores exhibited a significant negative correlation with the diffusion free energy profiles ($R^2=0.971$, $P < 0.0001$), whereas the correlation with the experimentally measured apparent permeability was relatively weak ($R^2=0.236$, $P < 0.5$), likely because of the out-of-distribution scaffold of the tested candidates. These results suggested that the tested alkaloids were highly likely to traverse the BBB through passive diffusion mechanism and that SynNat-BERT’s predictivity of real-world apparent permeability (Fig. 6c).

To elucidate specific structural characteristics associated with BBB permeability, we extracted key conformational states within diffusion pathway and analyzed the intermolecular interactions between tested alkaloids and DOPC lipids (Fig. 6d and Supplementary Fig. 9). Based on the measured P_{app} values and the predicted SynNat-BERT scores, we identified a hydrogenated berbine scaffold, tetrahydropalmatine, exhibiting higher permeability compared to other unhydrogenated scaffolds, such as palmatine and chelerythrine (2.01 ± 0.04) $\times 10^{-6}$ cm/s and 1.76 ± 1.02) $\times 10^{-7}$ cm/s, respectively, Supplementary Tab. 6). We visualized the atomic attention maps reported by SynNat-BERT to systematically investigate the potential differential biological activities (Supplementary Fig. 10). The attention mechanism highlighted that carbon atoms within the hydrogenated rings were important to penetration, whereas these atoms showed no significant importance within aromatic alkaloids, suggesting model’s interpretability of penetration mechanism (Fig. 6e). Intermolecular interactions also supported this observation by highlighting more π -alkyl interactions to DOPC carbon tails with flexible conformations in the tetrahydropalmatine scaffold. In contrast, rigid protoberberine scaffold in palmatine can only form intermolecular interactions with DOPC carbon tails that were parallel to the plane (Fig. 6e). In addition, we investigated molecules possessing similar scaffolds to validate this hypothesis. SynNat-BERT revealed consistently high prediction scores of berbine scaffold (0.618 ± 0.041), elucidating an activity cliff exists between hydrogenated and unhydrogenated scaffolds (see Supplementary Note. 2). Meanwhile, we simulated berbine-, protoberberine-, and benzo[c]phenanthridine scaffolds and found that aromatic alkaloid scaffolds, such as protoberberine and benzo[c]phenanthridine, were unfavorable for BBB penetration with positive energy barriers (Fig. 6f). These evidence further confirmed the penetration potential of the hydrogenated berbine scaffold via a passive diffusion mechanism. Overall, through proof-of-concept virtual screening, our model identified unique BBB-permeable compounds with a potent hydrogenated berbine scaffold. Moreover, the attention map and intermolecular interaction-facilitated interpretability analysis validated the underlying passive diffusion mechanism.

Discussion

In this study, we present SynNat-BERT, a substructure-based self-supervised learning framework that integrating dual-modal adaptation for the prediction and structural interpretation of small-molecule BBB permeability. Through comprehensive evaluations of embedded chemical space, predictive accuracy, and attention weight distributions, our results demonstrate the multifacet functionality of the cross-attention fusion architecture. Under the stringent scaffold-splitting condition, SynNat-BERT consistently outperformed existing baseline methods in classification accuracy and regression correlation performances.

One of the key contributions of this work lies in the efficient self-supervised pre-training enabled by a high-quality integrated data set and an interpretable substructure tokenizer. By incorporating 1.7 million unlabeled molecules from synthetic and natural product databases, SynNat-BERT effectively extracted generalizable molecular textual representations and constructed a natural product-aware chemical foundation model. Building upon this foundation, we develop the dual-modal adaptation framework employing a cross-attention fusion mechanism to address key challenges in quantitatively predicting the BBB permeability. Molecular text and physicochemical descriptors are encoded separately, followed by a Gaussian reweighting transformation to map heterogeneous molecular properties onto a unified scale, mitigating long-tail effects

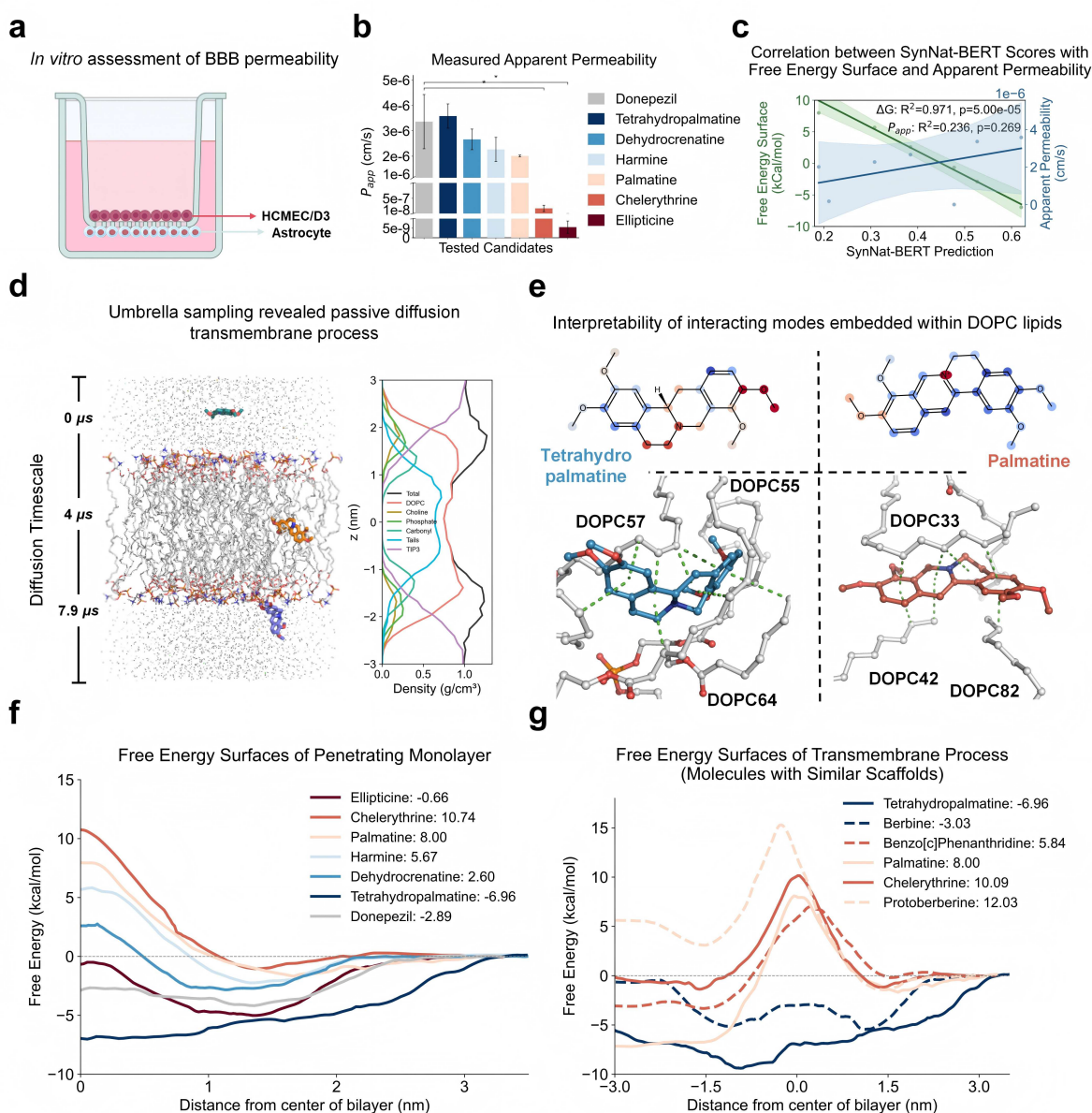


Figure 6: *In vitro* permeability assessment of SynNat-BERT identified candidate alkaloids. (a) Schematic illustration of the bilayer co-cultured Transwell system used in this study to assess the *in vitro* BBB permeability of screened candidates. (b) Measured apparent permeability (P_{app}) of the in-house tested molecules ($n=7$). Positive control, donepezil, is colored in gray. Errors derived from three replicate sampling for each molecule ($n=3$). (c) Correlation between predicted logBB and calculated free energy profiles. Light green area represents 95% confidence interval (CI). (d) Schematic diagram of molecular dynamics simulating passive diffusion across bilayer membrane system. Density distributions of system components including water (TIP3) and partial lipids are visualized in the right panel. (e) Interpretability analysis of hydrogenated and unhydrogenated palmatine. Atomic attention weights are shown in the above, while 3D interaction modes are exhibited below. The π -alkyl interaction is depicted as green dashed lines, and key interactions with DOPC indexes within 5 Å are annotated. (f) Free energy surfaces of the tested alkaloids compared to donepezil ($n=7$). (g) Free energy surfaces of the tested molecules with berberine scaffolds ($n=6$). Free energy surfaces of scaffolds are represented by dashed lines. Colors of the tested candidate correspond to panel b.

and improving cross-modal alignment. In addition, feature engineering validates a minimum input of seven interpretable physicochemical features to discover structurally distinct BBB-permeable candidate, offering a feasible paradigm for multi-modal learning in real-world molecular property prediction and novel scaffold hopping.

Despite its robustness, the study has certain limitations. One notable constraint is that our proof-of-concept screening only select a subset of candidates from alkaloid natural products for prospective experimental validation. This strategy enabled us to verify the model’s capability of discovering hits with novel scaffold. Indeed, we did not exclude the possibility that candidates may also exist in other natural product categories, such as terpenoids and polyketides. While the developed pipeline was assessed in the context of alkaloid natural products, chemical space can be extended to more valuable area of penetrable peptide discovery through a similar strategy as described. Meanwhile, as data availability continues to expand, the integration of currently measured *in vitro* permeability will facilitate high-quality development of quantitative data sets for BBBP from different experimental sources. Future studies could extend the framework by incorporating high-throughput organoid-derived data sets to achieve improved physiological relevance and enable more precise investigation of the molecular mechanism underlying BBB penetration.

In summary, we demonstrate that self-supervised learning frameworks integrating molecular context and physicochemical properties can be exploited to accurately predict and interpret molecular physiological properties, thus facilitating the identification of structurally distinct candidates from the vast natural product library. While this work focused on BBBP prediction as a proof-of-concept, the developed approach can easily be adapted to other application scenarios. By leveraging heterogeneous *in vivo* and *in vitro* data under different experimental conditions, we hope to suggest promising directions for future research on molecular property prediction and scaffold discovery.

Experimental

Pre-training data curation

To take molecular diversity into consideration, molecules from PubChem and COCONUT databases were utilized to form an unlabeled merged data sets, containing both chemical compounds and natural products. PubChem (10 million), a widely used library for pre-training chemical language models, contains information from hundreds of sources all over the world[58]. COCONUT is an open-source platform of depositing, curating and reusing natural products data, containing 700,000 molecules (COCONUT 2025-08)[59]. In this study, the pre-training model performance was benchmarked across PubChem and COCONUT data sets with varying sizes to investigate the impact of the data scale on model’s comprehension of chemical information. SMILES augmentation strategy was applied to alter the scale of chemical data, where duplicated molecules were dropped during enumeration[60]. Thus, SynNat-BERT-1.7M and SynNat-BERT-10.7M were pre-trained on a combination of (100% COCONUT + 10% PubChem, 1.7M) and (100% COCONUT + 100% PubChem, 11M), while we also provided model variants (28×100% COCONUT + 100% PubChem, 30M) to study the effect of increasing the scale of pre-training data. In addition, model purely pre-training on COCONUT were benchmarked to understand the impact of combined sources of data sets, thereby creating model variants (2×100% COCONUT, 1.4M) small model, (12×100% COCONUT, 8.4M) intermediate model, and (28×100% COCONUT, 20M) large model (Supplementary Fig. 1).

Post-training data processing

Within post-training classification adaptation, we first collected molecular samples with qualitative BBB permeability data from MoleculeNet data set[17]. Data preprocessing was performed to achieve fair comparisons across benchmarking models: (1) invalid SMILES with erroneous molecular grammar were filtered out; (2) molecules with missing physicochemical properties were given ‘nan’ values. To perform more rigorous evaluations, data sets were split based on Bemis-Murcko scaffold, where structurally similar compounds were grouped within identical subsets, ensuring testing sets to contain unseen chemical scaffolds[20].

For regression task, 1,058 molecules with quantitative labels were collected from B3DB, currently the largest BBBP database with realistic experimental log BB values[18]. Both classification and regression post-

training data sets were structurally split into training (80%), testing (10%), and validation (10%) sets with five-fold cross-validation to alleviate scaffold splitting bias.

Pre-training model architecture

The original Bidirectional Encoder Representations from Transformers (BERT) model is defined as an encoder-only transformer, which contains three major components, a positional embedding layer, several transformer encoder layers, as well as a task-specific prediction layer for downstream adaptation. Inspired by previous BERT architectures applied for molecular representation learning and property predictions, we adopted a similar strategy utilizing large-scale unlabeled molecular data to pre-train SynNat-BERT framework based on RoBERTa model implementation in an unsupervised pattern[34]. In this context, each SMILES was analogous to a 'sentence' composed of multiple words and characters. In the embedding layer, input SMILES sequences were first segmented into substructural vectors together with special tokens (e.g. [CLS] and [SEP]) based on the predefined byte-pair encoding (BPE) vocabulary. Specifically, we defined the pre-training data structure as a tensor $X \in \mathbb{R}^{N \times 2D_t}$, where N was the number of molecules in the pre-training molecular library, H was the number of attention heads, and D was the embedding dimension. For each molecule i, the input consisted of two component: (1) token embeddings ($\in \mathbb{R}^{1 \times D_t}$), represented by dense vectors of token j, and (2) multi-head self-attention masks ($\in \mathbb{R}^{1 \times D_t}$), encoded by attention score of each token in the input sequence for all other tokens in the self-attention fashion. Each tokenizer vector was padded into a fixed-length embedding vector of dimension $D_t = 512$. Combined with the positional encoding vector, the two components were concatenated to represent the final input for molecule i ($X_i \in \mathbb{R}^{2 \times D_t}$).

Token embedding

The input of tokens $x_j = (x_1, x_2, \dots, x_{D_t})$ was encoded based on substructural vocabulary ($\in \mathbb{R}^V$), where V was the vocabulary size. In this study, the maximum lengths of BPE and ST (SMILES-Tokenizer) vocabulary were defined as 7,924 and 591, respectively, determined by all the unique tokenizers encountered during ChemBERTa pre-training phase[61]. Each element x_j was mapped to the corresponding unique integer ID in the vocabulary. Special tokens such as masking (<mask> for BPE and [MASK] for ST), beginning of sentence (<s> for BPE and [CLS] for ST), end of sentence (</s> for BPE and [CLS] for ST), padding (<pad> for BPE and [PAD] for ST), etc., were excluded from embedding to ensure these tokens were not split during segmentation. The token embedding for a specific molecule $T_{M_i} \in \mathbb{R}^{D_t}$ can thus be expressed as:

$$T_{M_i} = [T_{\text{cls}}, T_1, T_2, \dots, T_D, T_{\text{sep}}], \quad (1)$$

where T_j represented the specific token ID decided by the corresponding position in tokenizer vocabulary.

Attention mask embedding

The attention weight of each input sequence was calculated via multi-head self-attention mechanism. To determine whether specific position was valid to be involved in attention calculation or was padding token, attention mask vector ($A_{M_i} \in \mathbb{R}^{D_t}$), a binary tensor, was generated during SMILES segmentation. Thus, it was able to control which positions were trainable to the model. The final input embedding of a specific molecule ($X_i \in \mathbb{R}^{D_t}$) was constructed by:

$$X_i = T_{M_i} + A_{M_i} \quad (2)$$

The self-attention mechanism implemented in this study is similar to the Transformer encoder is named scaled dot-product attention, which maps the input data X into H distinct query matrix $Q_h = XW_h^Q$, key matrix $K_h = XW_h^K$, and value matrix $V_h = XW_h^V$ [62], where $h = 1, 2, \dots, H$ is the number of attention head, W_h^Q , W_h^K , and W_h^V are learnable parameters. After linear projections, the scaled dot-product attention is formulated as:

$$O_h = \text{Softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) V_h, \quad (3)$$

where d_k is a scaling factor and O_h is the output of a specific head in the attention layer. The output of different attention heads ($O_1, O_2, \dots, O_H$) are then concatenated and go through multi-layer perception (MLP) to mix up information. Upon the multi-head self-attention output, the resulting hidden states are further transformed by a linear projection, a Gaussian error linear unit (GELU) activation function, followed by a layer normalization to increase generalizability. The last linear decoder maps the hidden representations to prediction probability of each token towards the vocabulary, thus realizes the masked multi-classification task. The Transformer encoder layer was shared between pre-training and post-training stages.

Masked language modeling pre-training objective

In this study, we selected masked language modeling (MLM) as the pre-training strategy, where 15% of substructural tokens were randomly masked. Model was trained to correctly identify them to learn generalized molecular language representations. Special tokens (e.g. [CLS], [PAD]) were excluded from masking. To ensure SynNat-BERT captured generalized molecular context, we conducted a thorough hyperparameter search varying the number of hidden layers, number of attention heads, dropout, size of vocabulary, and size of intermediate layer. Hyperparameters were determined by evaluating molecular recovery accuracy and visualizing distributions of unlabeled molecules from pre-training library within low-dimensional latent space. Pre-trained with AdamW optimizer for 3 epochs, the optimal model contained 6 attention heads and 6 hidden layers with 64 hidden neurons per head. The initial learning rate was set to 5×10^{-5} , and the batch size was set to 32. This configuration yielded around 63.7 million parameters.

Post-training model architecture

In addition to the above pre-training input embedding, we introduced an auxiliary descriptor modality to the overall input in the post-training stage to integrate molecular contextual information and physicochemical features. We systematically investigated the prediction accuracy of the permutation and combination of adding various physicochemical features (see Supplementary Note. 1). The optimal adaptation model contained 7 well-established interpretable features ($D_d = 7$), including: (1) hydrogen bonding (NumHydrogenBondDonor, and NumHydrogenBondAcceptor)[41]; (2) lipophilicity (MolLogP)[42]; (3) molecular size and volume (MolWt)[43]; (4) molecular heteroatoms (NOCOUNT)[44, 45]; (5) molecular shape and flexibility (NumRotatableBonds)[46]; (6) molecular topology (TPSA)[47]. Before performing transformation, we first handled the missing values (NaN) in the descriptor data. For each feature column $f \in 1, 2, \dots, D_d$, median number m_f was calculated by:

$$m_f = \text{median}(\{d_i^f \mid d_i^f \text{ is not NaN, } i = 1, 2, \dots, N\}), \quad (4)$$

where d_{if} was feature value f of molecule i . When d_{if} was NaN, it was replaced by the median value of that column m_f .

Descriptor embedding

To enable consistency between each physicochemical property, we introduced a two-stage transformation: (1) Gaussian reweighting using a quantile transformation mapped the empirical distribution of a descriptor to a standard normal distribution; (2) PCA (principal component analysis) was then applied to the Gaussianized feature space to filter decorrelated features.

For each f in feature column, calculated descriptor was transformed based on the quantile ranking among all molecules. The calculated descriptor matrix of the training set $\mathbf{d}_{\text{train}} \in \mathbb{R}^{D_d}$ was firstly encoded and fitted into a quantile transformer using empirical cumulative distribution function (ECDF, F_f):

$$d_f^G = \Phi^{-1}(F_f(d_f)), \quad (5)$$

where d_f was the feature f quantile array for all molecules, $\mathbf{d}_f^G \sim \mathcal{N}(0, 1)$ was the Gaussian reweighted feature matrix for descriptor f , and Φ^{-1} was the inverse cumulative distribution function (CDF) of the standard normal distribution.

Then, principal component analysis was performed on $\mathbf{d}^G$:

$$\mathbf{d} = \bar{\mathbf{d}}^G \mathbf{V} \mathbf{\Lambda}^{-1/2}, \quad (6)$$

where $\bar{\mathbf{d}}^G$ was the centralized matrix based on the mean value of $\mathbf{d}^G$, $\mathbf{V}$ was the covariance matrix eigenvector, and $\mathbf{\Lambda}^{-1/2}$ was the eigenvalue matrix. Then, calculated feature array in the testing and validation set ($\mathbf{d}_{test}$ and $\mathbf{d}_{valid}$, respectively) were transformed accordingly based on the fitted training transformer.

Cross-attention fusion

In this work, we applied cross-attention fusion mechanism to integrate different modalities. This operation allowed the model to dynamically weight physicochemical features conditioned on molecular semantic context. The descriptor matrix $\mathbf{d}$ was projected onto the same representation space ($\in \mathbb{R}^{N \times D_d}$) of textual embedding via a linear transformation:

$$\mathbf{d}' = \mathbf{W}_d \mathbf{d} + b, \quad (7)$$

where $\mathbf{W}_d \in \mathbb{R}^{D_d \times D_t}$ was the weight matrix of descriptor modality, and b was the trainable bias learned through back propagation. Subsequently, multi-head cross-attention module with H attention heads was applied, using the textual embedding as the query and the project descriptor embedding as key and value:

$$\mathbf{q} = \mathbf{t} \mathbf{W}_q, \mathbf{k} = \mathbf{d}' \mathbf{W}_k, \mathbf{v} = \mathbf{d}' \mathbf{W}_v, \quad (8)$$

where weight matrices $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{D_t \times D_t}$. For h th single attention head, the attention weight O_h was expressed as:

$$\mathbf{O}_h = \text{Softmax}\left(\frac{\mathbf{q}_h \mathbf{k}_h^T}{\sqrt{d_k}}\right) \mathbf{v}_h, \quad (9)$$

where $\sqrt{d_k}$ is the dimension of each head h , which equaled to D_t/h . The multi-head attention was then concatenated and passed through linear projection:

$$\mathbf{O} = \text{Concat}(\mathbf{O}_1, \dots, \mathbf{O}_H) \mathbf{W}_O, \quad (10)$$

where weight matrices $\mathbf{W}_O \in \mathbb{R}^{D_t \times D_t}$. In this work, we employ $H = 6$ parallel attention heads. The output of the cross-attention module was further processed by an MLP to produce a fused molecular representation:

$$\mathbf{f} = \text{MLP}_{\text{fusion}}(\mathbf{O}) = \text{GELU}_2(W_2(\text{Dropout}(\text{GELU}_1(W_1 \text{LayerNorm}(\mathbf{O})))))) \quad (11)$$

Classifier head

During post-training adaptation, the pre-trained text encoder was frozen, while the fusion and classification layers were trained to incorporate descriptor modality. The fused representations were passed through a BBBP task-specific prediction head to replace pre-training head. The encoded hidden layer of BERT for the input molecules X was expressed as:

$$\mathbf{H} = \text{Encoder}(X) \quad (12)$$

Mean pooling was applied to extract text vector $\mathbf{t}$. Thus, the final output can be simplified by a cross-attention multi-head attention (MHA) mechanism:

$$\mathbf{f} = \text{MHA}(\mathbf{t}, \mathbf{d}', \mathbf{d}') \quad (13)$$

The classifier head used the fused representations to predict BBBP labels by an MLP implementing rectified linear unit (ReLU) activation function:

$$y_i = \text{MLP}_{\text{cls}}(\mathbf{f}) = \mathbf{W}_2(\text{Dropout}(\text{ReLU}(\mathbf{W}_1 \text{LayerNorm}(\mathbf{f})))) \quad (14)$$

For BBBP binary classification task, cross-entropy loss was minimized:

$$\mathcal{L} = - \sum_i [\hat{y}_i \log y_i + (1 - \hat{y}_i) \log(1 - y_i)], \quad (15)$$

where $\hat{y}_i$ was the true label vectors in different batches i , and y_i was the probability logits calculated from MLP_{cls}. Detailed hyperparameter settings of SynNat-BERT see Supplementary Tab. 7.

Prospective screening protocols

In this study, the proof-of-concept screening was performed using the pre-trained and adapted SynNat-BERT-1.7M (100% COCONUT + 10% PubChem) model under the entire pre-trained chemical space (100% COCONUT + 10% PubChem, ~ 1.7 M) to screen unseen molecules without enumeration. To concentrate on BBB permeable natural product categories, we classified natural products based on 'np_classifier_pathway' labels given in the COCONUT library into terpenoids, alkaloids, shikimates and phenylpropanoids, fatty acids, polyketides, amino acids and peptides, and unknown[59]. To narrow into representative BBB permeable hits with CNS penetrant scaffolds, we excluded potential scaffolds with (1) too simplified backbones with single or double aromatic rings, such as benzene, pyridine and indole, etc.; (2) backbones without nitrogen atoms that was not representative alkaloids; (3) backbones with at least 5 molecules containing the same scaffold.

Implementation details

The SynNat-BERT model was implemented using PyTorch (v2.9.1) and Transformers (v4.36.2) as the deep learning backend[63, 64]. Descriptor preprocessing were calculated by using RDkit (2024.09)[65]. Gaussian quantile and PCA transformations were implemented with scikit-learn package (v1.7.2)[66]. Key dependencies included safetensors (v0.5.3), scipy (v1.15.3), umap (v0.5.9.post2), and TensorFlow (v2.19.0), etc. The model was trained on NVIDIA RTX-4090 GPU (24GB) to achieve pre-training and post-training consistency. Pre-training on 10.7M dataset took approximately 5 h, while pre-training on 1.7M took around 3 h. Post-training on MoleculeNet dataset (n=2,039) took around 1.5 h, while post-training on B3DB regression dataset (n=1,058) took approximately 0.5 h. Detailed hyperparameter settings for baseline models see Supplementary Tabs. 8-10.

Evaluation metrics

To facilitate a fair comparison of SynNat-BERT with other baselines and model variants, we computed a total of eight metrics, classified into three categories: accuracy metrics (for example, recovery accuracy (r.a.) and AUROC), attention metrics (for example, mean cosine similarity (m.c.s.)), correlation metrics (for example, PCC). These metrics provided a thorough overview of the model’s comprehensive performance in prediction accuracy, regression correlation, and screening capability.

Accuracy metrics

Recovery accuracy (r.a.): r.a. measures the the fraction of masked token for which the predicted token exactly matches the ground-truth token. The formula is expressed by:

$$\text{r.a.} = \frac{1}{|M|} \sum_{i \in M} \mathbb{I}(y_i = \hat{y}_i), \quad (16)$$

where M denotes the set of masked tokens, y_i is the predicted token at position i , $\hat{y}_i$ is the ground-truth token at position i , and $(\cdot)$ is the indicator function which equals to 1 if $y_i = \hat{y}_i$.

Area under receiver operating curve (AUROC): AUROC computes the probability that a randomly selected positive sample has a predicted score high than that of a randomly selected negative sample. The formulation can be denoted as:

$$\text{AUROC} = P(s(x^+) > s(x^-)) = \int_0^1 \text{TPR}(\text{FPR}) d(\text{FPR}), \quad (17)$$

where $s(x^+)$ and $s(x^-)$ represent the predicted score of a randomly selected positive and negative samples, respectively. FPR and TPR denote false positive rate and true positive rate, respectively:

$$\text{FPR} = \frac{FP}{FP + TN}, \text{TPR} = \frac{TP}{TP + FN} \quad (18)$$

Area under precision-recall curve (AUPRC): AUPRC assesses the trustworthiness among positive predictions, which can be formulated as follow:

$$\text{AUPRC} = \int_0^1 \text{Precision}(\text{Recall}) d(\text{Recall}) \quad (19)$$

where Recall can be calculated by TPR, while Precision is expressed as:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (20)$$

Attention metrics

Mean cosine similarity (m.c.s): m.c.s. calculated the overall similarity of attention patterns across different attention heads H :

$$\text{m.c.s.} = \frac{1}{H(H-1)} \sum_{u \neq v} \frac{\mathbf{a}_u \cdot \mathbf{a}_v}{\|\mathbf{a}_u\| \|\mathbf{a}_v\|}, \quad (21)$$

where a_u and a_v are the attention weight outputted from attention head u and v , respectively, and H is the total number of attention heads.

Correlation metrics

Pearson correlation coefficient (PCC): PCC measures the linear relationship between calculated physicochemical features y_i and BBBP labels y_i of molecule i within MoleculeNet data set. The formula can be expressed by:

$$\text{PCC} = \frac{\sum_{i=1}^N (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2}}, \quad (22)$$

where $\bar{y}$ and $\bar{\hat{y}}$ are the means of predicted and ground-truth values. In this study, $|r_{\text{BBB}}|$ was used to represent physicochemical features correlations to BBBP labels, while $|r_{\text{feat}}|$ was applied to measure the inter-correlation between each incorporated feature.

Molecular dynamics simulations of membrane permeability

Molecular dynamics (MD) have been proven to be effective to predict blood-brain barrier (BBB) permeability of small molecule compounds in a physics-based pattern[67]. In this study, we followed previous reported method to apply umbrella sampling technique to recover natural products’ free energy profiles of passive diffusion to represent the spontaneity to pass through BBB[52]. The following potential BBB permeable alkaloids were chosen to run MD simulations: tetrahydropalmatine, palmatine, dehydrocrenatine, harmine, chelerythrine, ellipticine and donepezil. Alkaloid scaffolds berbine, protoberberine, benzo[c]phenanthridine were also simulated to compare the permeability. These alkaloids molecules were selected from SynNat-BERT predicted BBB+ set based on proof-of-concept screening with a scaffold-similar CNS-penetrable drug, donepezil, which was capable to cross BBB.

MD setup

All simulations were run using GROMACS-2024.2[68]. Charmm36m force field was used for 1,2-dioleoyl-sn-glycero-3-phosphocholine (DOPC) molecules[69], CGenFF force field was used for all small molecules[70], and the TIP3P model was applied to represent water molecules[71]. Each simulation was performed in a rectangular lipid bilayer system ($5 \times 5 \times 8.5 \text{ nm}^3$) containing a single copy of the molecule. The bilayer membrane system was constructed using CHARMM-GUI server [<https://www.charmm-gui.org/?doc=input/membrane.bilayer>][72]. A typical simulation system contained ~ 3240 water molecules, 72 DOPC, 7 Na^+ , 7 Cl^- , and one small molecule to form a 0.15 M neutralized system with around 19,700 atoms in total.

Simulation systems were first subjected to a 50,000 step energy minimization using the steepest descent algorithm, followed by two 125 ps temperature equilibration runs at 303.15 K using a velocity rescaling thermostat[73]. Subsequently, the system underwent a 125 ps and three 500 ps NPT equilibrations at 1 atm bar constant pressure with a semi-isotropic rescaling barostat ($\tau_p = 1$, compressibility= $4.5 \times 10^{-5} \text{ bar}^{-1}$), maintaining the temperature at 303.15 K[74]. Throughout the equilibration stages, positional constraints on the protein backbone, side-chain, and lipid phosphorus atoms were gradually reduced to ensure system stability before the production run. Van der Waals interactions were treated with a 12 Å cutoff, and long-range coulomb interactions were calculated using the particle mesh Ewald (PME) method[75]. Non-bonded interactions were handled with buffered neighbor searching via the Verlet algorithm[76], while hydrogen bond lengths were constrained using the LINCS algorithm[77].

The topologies of small molecules were modeled using CHARMM-GUI server [<https://www.charmm-gui.org/?doc=input/ligandrm>]. After building topologies, quantum mechanical calculations were used to produce parameters that were close to physical conditions. Initially, molecules were pre-optimized at the HF/STO-3G level of theory, and then reoptimized using B3LYP/6-31G density functional theory (DFT) with implicit water setting as polarizable continuum model (PCM)[78].

Free energy calculations

Umbrella sampling was used to calculate the potential of mean force (PMF) free-energy surface for the passive diffusion of molecules. A harmonic restraint with a force constant of $1,000 \text{ kJ} \cdot \text{mol}^{-1} \cdot \text{nm}^{-2}$ along z axis (the normal to the bilayer) was applied to the distance between the center of mass of the DOPC bilayer and the center of mass of the molecule. The pulling simulations lasted for around 1 ns on average to obtain a complete trajectory of a molecule to pass through the membrane. ~ 900 umbrella sampling were performed in increments of 1 ps along the z axis direction to ensure thorough sampling to convergence. The center of the bilayer was treated as $z=0$, thus the window spanned $z=-3.5 \text{ nm}$ to $z=3.5 \text{ nm}$. The 7 nm distance completely covered the diffusion of molecules from bulk water, through the entire bilayer membrane, and then back into bulk water again.

Each simulation window was run for 10 ns, for a total of $\sim 63 \mu\text{s}$ simulation time for the 7 different molecules. The initial of each simulation was regarded as extensive equilibration time for the compound to adjust to a correct orientation to pass through the membrane. The weighted histogram analysis method (WHAM) was used to calculate the PMF for each compound transforming biased sampling to unbiased distributions[79]. The overlap between neighboring replicas was examined to check whether the PMF curves had converged. If needed, simulations were extended and parameters are adjusted to achieve convergence. All PMF analysis were carried out on the $[0, 3.5] \text{ nm}$ regions along z axis, since the lipid bilayer can be treated as a symmetrical system. All free-energy surface were then normalized and aligned to zero in bulk water.

In vitro Transwell assay to determine BBB permeability

Reagents and chemicals

Cell Culture Insert 12-Well Plate Set ($0.4 \mu\text{m}$ Pore Size, PET Membrane) was purchased from Beyotime Biotech Inc. Dehydrocresatinine (26585-13-7) was obtained from Beijing Solarbio Science & Technology Co., Ltd. Palmatine (3486-67-7), Cephaeline (Internal Standard, IS) (483-17-0) and Harmine (442-51-3) were purchased from Shanghai yuanye Bio-Technology Co., Ltd. Tetrahydropalmatine (6024-85-7) was purchased

from Shanghai Hongyong Biotechnology Co., Ltd. Donepezil (120014-06-4) was purchased from MedChem-Express. Matrigel® Basement Membrane Matrix was purchased from Sigma-Aldrich.

Preparation of standard samples

In the BBB cell model, each standard stock solution (Dehydrocrenatine, Palmatine, Harmine, Tetrahydropalmatine at 50 mM) was dissolved in dimethyl sulfoxide (DMSO) and Donepezil at 5.27 mM was dissolved in methanol. For cytotoxicity evaluation, the stock solutions were diluted to multiple concentration gradients using ECM medium. In the LC-MS experiment, stock solutions were prepared at 1 mg/mL in methanol, and working standard solutions were prepared by diluting the stock solutions with methanol.

Cytotoxicity evaluation

The cytotoxicity of four alkaloid compounds and the positive drug donepezil was evaluated using HCMEC/D3 and AST cells. HCMEC/D3 and AST cells were seeded into 96-well culture plates at a density of 8×10^3 cells per well. After culturing in ECM supplemented with 10% fetal bovine serum for 24 hours, the cells were treated with Dehydrocrenatine, Palmatine, and Tetrahydropalmatine (0, 5, 10, 25, 50, and 100 μ M), Donepezil (0, 8, 16, 32, 64, 128 μ M) and Harmaline, Ellipticine, Chelerythrine (0, 1, 2, 3, 4, and 5 μ M) for 24 hours. Subsequently, cell viability was assessed using the CCK-8 assay.

Co-culture of rat HCMEC/D3 and astrocytes as BBB cell model

Transwell polycarbonate inserts (surface area 0.6 cm²; pore size 0.4 μ m) were coated with Matrigel prior to use. Firstly, the prepared astrocytes were seeded onto one side of the inserts at a density of 8×10^4 cells per insert. After allowing the astrocytes to attach to the insert for 1.5 hours, HCMEC/D3 cells were seeded onto the opposite side of the inserts at 1×10^5 cells per insert. Finally, in vitro experiments were conducted three to four days after HCMEC/D3 cell seeding.

Transport experiment in Human BBB cell model

Transport experiments were initiated by adding alkaloid compounds at final concentrations to the apical side (AP, 0.4 mL), with AM medium added to the basolateral side (BL, 0.4 mL). To determine the time-dependent transport of alkaloid compounds, aliquots (60 μ L) were collected from the BL at 15, 30, 60, 90, 120, 240, and 360 min, and immediately replaced with an equal volume of fresh AM medium. An aliquot (60 μ L) was collected from the BL side at 60 min to determine the P_{app} values of the alkaloid compounds.

Instrumentation and LC-MS assays of alkaloids

Quantitative analysis was performed using a UHPLC system (Shimadzu) equipped with an API 4500 triple quadrupole mass spectrometer (AB SCIEX, Concord, Ontario, Canada) with an ESI source. An ACQUITY UPLC Ben C18 chromatographic column (1.7 μ m, 2.1 $\times$ 100 mm) was used to separate the compounds. The column temperature was maintained at 40 °C, the flow rate was 0.3 mL/min, and the injection volume was 2 μ L. The mobile phase consisted of 0.1% formic acid in water (A) and acetonitrile (B). The gradient elution program was as follows: 0-0.1 min, 50% B; 0.1-4 min, 50-95% B; 4-5 min, 95-95% B; 5-6 min, 95-5% B; 6-6.5 min, 5-50% B. The relevant mass spectrometry parameters were as follows: curtain gas (CUR): 30 psi; collision gas (CAD): 9 psi; ion spray voltage (IS): -4500 V / +5500 V; ion source temperature (TEM): 400 °C; nebulizer gas (GS1): 50 psi; turbo heater gas (GS2): 50 psi.

Sample preparation

BL sample (30 μ L) was mixed with 30 μ L of methanol and 30 μ L of IS (100 ng/mL). The mixture was then vortex-mixed with 120 μ L of acetonitrile for 5 min. Following centrifugation at 14,000 rpm for 15 min, the supernatant was transferred to a clean tube and evaporated to dryness under a gentle stream of nitrogen. The resulting residue was reconstituted in 30 μ L of methanol, vortexed for 5 min, and centrifuged again at 14,000 rpm for 15 min. Finally, the supernatant was injected into the UPLC-MS/MS system for analysis.

Apparent permeability calculation

In this study, we used the calculation equation of apparent permeability (P_{app}) proposed by [80].

$$P_{app} = \frac{dQ/dt}{SA \cdot C_0} \quad (23)$$

where dQ/dt represents the amount of substance transported across the membrane per unit time, SA denotes the membrane surface area of Transwell plate, while C_0 represents the initial concentration from the top channel.

Statistics and reproducibility

For model recovery performance evaluations, the paired t -test was used to compare the differences between SynNat-BERT pre-trained on the merged PubChem and COCONUT data set and other pre-training variants or RoBERTa baseline models (two-sided) under warm-start, cold-data and cold-train scenarios. For the comparison of predicted scores of natural product categories, the pairwise independent Welch’s t -test was performed to compare the predicted scores of the reference group (alkaloid) to others. For the comparison of predicted scores of different alkaloid scaffolds, the pairwise independent Welch’s t -test was performed to assess the significance of the differences between three alkaloid scaffolds. The significance level was set as **** $P \leq 0.0001$; *** $P \leq 0.001$; 0.001 $\leq$ ** $P \leq 0.01$; 0.01 $\leq$ * $P \leq 0.05$; n.s., $P \geq 0.05$.

List of abbreviations

AttentiveFP, attentive fingerprint; BBB, blood-brain barrier; BBBP, blood-brain barrier permeability; BERT, bidirectional encoder representations from transformers; BPE, byte-pair encoding; CNS, central nervous system; CV, cross-validation; D-MPNN, directed message-passing neural network; MD, molecular dynamics; MLM, masked language modeling; PCC, pearson correlation coefficient; PMF, potential of mean force; QSAR, quantitative structure-activity relationship; RF, random forest; SMILES, simplified molecular-input line-entry system; SSL, self-supervised learning; TPSA, topological polar surface area; UMAP, uniform manifold approximation and projection; XGBoost, extreme gradient boosting.

Availability of data and materials

For pre-training data set, we obtained 695,114 natural products from the COCONUT library (<https://coconut.naturalproducts.net/>), 1,000,000 chemical compounds from the PubChem database (<https://pubchem.ncbi.nlm.nih.gov/>), 2,854,815 compounds from the ChEMBL v36 database (https://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_36/) and 249,456 compounds downloaded from ZINC data set (<https://www.kaggle.com/datasets/basu369victor/zinc250k>). For post-training adaptation, raw data was obtained through MoleculeNet data set [17] (<https://moleculenet.org/datasets-1>). Regression models were trained on 1,058 molecules with logBB values based on the B3DB data set [18] (https://github.com/theochem/B3DB/blob/main/B3DB/B3DB_regression.tsv). For interpretation of CNS penetrant scaffolds, we obtained 812 CNS permeable compounds from SelleckChem (<https://www.selleckchem.com/contact.html>).

The source code of SynNat-BERT is provided as open-source software under MIT lincense in the following repository: <https://github.com/ddddd-junlin/SynNat-BERT.git>. All conclusions present in the paper can be reproduced using the provided scripts.

Competing interests

The authors declare no competing financial interest.

Funding

This work was supported by the National Natural Science Foundation of China (82104357), Shenzhen Science and Technology Program (KJZD20240903102714019), and Shenzhen Science and Technology Program (QNXMC20250701101259014).

Author Contributions

† J.D., S.H., and S.L. contributed equally to this work. X.Y., Y.T., and B.W. supervised the project. J.D., S.H., and D.Z. contributed to the algorithm development and implementation. S.L., S.C., Z.Z., and Y.J. contributed to the in vitro experiment. J.D., S.H., S.L., and S.C. wrote the manuscript. Y.S., V.H., X.Y., Y.T., and B.W. revised the manuscript. All authors have given approval to the final version of the manuscript.

Acknowledgements

We want to thank Hailu Yan and Yingru Chen for training the model and providing analytical tools.

Supporting information

The Supporting Information is available free of charge. Supplementary figures (Figure S1-S15), raw data of ablation results (Table S1-S3), extended benchmarking results (Table S4-S5), detailed hyperparameters of all models (Table S7-S10), additional computational experiments of embedded features (Note S1, Table S11-S12), and additional computational validations of activity cliffs (Note S2). Additional experimental details (Table S6 and Figure S8) (PDF).

References

- [1] Terstappen, G. C.; Meyer, A. H.; Bell, R. D.; Zhang, W. Strategies for delivering therapeutics across the blood–brain barrier. *Nature Reviews Drug Discovery* 2021, 20, 362–383.
- [2] Vieira, D. B.; Gamarra, L. F. Getting into the brain: liposome-based strategies for effective drug delivery across the blood–brain barrier. *International Journal of Nanomedicine* 2016, 5381–5414.
- [3] Rawal, S. U.; Patel, M. M.; Patel, B. M. New drug delivery systems developed for brain targeting. *Drugs* 2022, 82, 749–792.
- [4] Profaci, C. P.; Munji, R. N.; Pulido, R. S.; Daneman, R. The blood–brain barrier in health and disease: Important unanswered questions. *Journal of Experimental Medicine* 2020, 217, DOI: 10.1084/jem.20190062.
- [5] Banks, W. A. From blood–brain barrier to blood–brain interface: new opportunities for CNS drug delivery. *Nature Reviews Drug Discovery* 2016, 15, 275–292.
- [6] Bickel, U. How to measure drug transport across the blood-brain barrier. *NeuroRx* 2005, 2, 15–26.
- [7] Chen, C.; Zhou, H.; Guan, C.; Zhang, H.; Li, Y.; et al. Applicability of free drug hypothesis to drugs with good membrane permeability that are not efflux transporter substrates: a microdialysis study in rats. *Pharmacology Research & Perspectives* 2020, 8, e00575.
- [8] Radan, M.; Djikic, T.; Obradovic, D.; Nikolic, K. Application of in vitro PAMPA technique and in silico computational methods for blood-brain barrier permeability prediction of novel CNS drug candidates. *European Journal of Pharmaceutical Sciences* 2022, 168, 106056.

- [9] Taillardat-Bertschinger, A.; Carrupt, P.-A.; Barbato, F.; Testa, B. Immobilized artificial membrane HPLC in drug research. *Journal of Medicinal Chemistry* 2003, 46, 655–665.
- [10] Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of Computer-Aided Molecular Design* 2013, 27, 675–679.
- [11] Fu, L.; Shi, S.; Yi, J.; Wang, N.; He, Y.; 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* 2024, 52, W422–W431.
- [12] Dobbelaere, M. R.; Lengyel, I.; Stevens, C. V.; Van Geem, K. M. Geometric deep learning for molecular property predictions with chemical accuracy across chemical space. *Journal of Cheminformatics* 2024, 16, 99.
- [13] Zhang, X.; Liu, T.; Fan, X.; Ai, N. In silico modeling on ADME properties of natural products: Classification models for blood-brain barrier permeability, its application to traditional Chinese medicine and in vitro experimental validation. *Journal of Molecular Graphics and Modelling* 2017, 75, 347–354.
- [14] Wang, Z.; Yang, H.; Wu, Z.; Wang, T.; Li, W.; Tang, Y.; Liu, G. In silico prediction of blood–brain barrier permeability of compounds by machine learning and resampling methods. *ChemMedChem* 2018, 13, 2189–2201.
- [15] Nguyen-Vo, T.-H.; Do, T. T. T.; Nguyen, B. P. An effective ensemble deep learning framework for blood-brain barrier permeability prediction. *2024 IEEE Conference on Artificial Intelligence (CAI)* 2024, 164–169.
- [16] Wang, Y.; Wang, Y.; Li, C.; Ji, W. A multi-view deep learning method for predicting blood-brain barrier permeability of peptides. *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2024, 7013–7020.
- [17] Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* 2018, 9, 513–530.
- [18] Meng, F.; Xi, Y.; Huang, J.; Ayers, P. W. A curated diverse molecular database of blood-brain barrier permeability with chemical descriptors. *Scientific Data* 2021, 8, 289.
- [19] Lachance, H.; Wetzel, S.; Kumar, K.; Waldmann, H. Charting, navigating, and populating natural product chemical space for drug discovery. *Journal of Medicinal Chemistry* 2012, 55, 5989–6001.
- [20] Scalia, G.; Rutherford, S. T.; Lu, Z.; Buchholz, K. R.; Skelton, N.; et al. Deep-learning-based virtual screening of antibacterial compounds. *Nature Biotechnology* 2025, 1–14.
- [21] Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence* 2022, 4, 1256–1264.
- [22] Qiang, B.; Lai, J.; Jin, H.; Zhang, L.; Liu, Z. Target prediction model for natural products using transfer learning. *International Journal of Molecular Sciences* 2021, 22, 4632.
- [23] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 2019, 4171–4186.
- [24] Zheng, X.; Tomiura, Y. A BERT-based pretraining model for extracting molecular structural information from a SMILES sequence. *Journal of Cheminformatics* 2024, 16, 71.
- [25] Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* 2022, 4, 279–287.

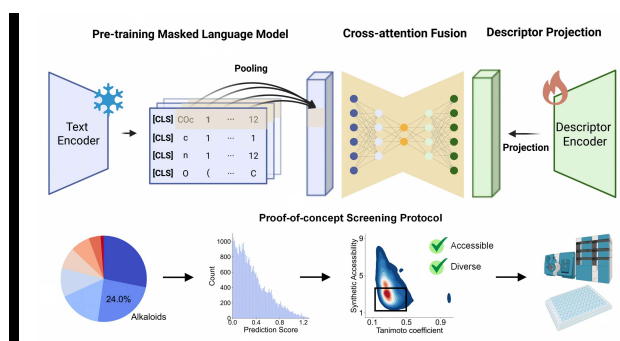
- [26] Lou, X.; Cai, J.; Siu, S. W. Multi-MoleScale: a multi-scale approach for molecular property prediction with graph contrastive and sequence learning. *Journal of Cheminformatics* 2025, 18, 5.
- [27] Koutroumpa, N.-M.; Tsoumanis, A.; Sarimveis, H.; Lynch, I.; Melagraki, G.; Afantitis, A. Prediction of blood–brain barrier and Caco-2 permeability through the Enalos Cloud Platform: combining contrastive learning and atom-attention message passing neural networks. *Journal of Cheminformatics* 2025, 17, 68.
- [28] Chilingaryan, G.; Tamoyan, H.; Tevosyan, A.; Babayan, N.; Hambardzumyan, K.; et al. BartSmiles: Generative masked language models for molecular representations. *Journal of Chemical Information and Modeling* 2024, 64, 5832–5843.
- [29] Zhang, X.-C.; Wu, C.-K.; Yi, J.-C.; Zeng, X.-X.; Yang, C.-Q.; Lu, A.-P.; Hou, T.-J.; Cao, D.-S. Pushing the boundaries of molecular property prediction for drug discovery with multitask learning BERT enhanced by SMILES enumeration. *Research* 2022, 2022, 0004.
- [30] Li, X.; Fourches, D. SMILES pair encoding: a data-driven substructure tokenization algorithm for deep learning. *Journal of Chemical Information and Modeling* 2021, 61, 1560–1569.
- [31] Li, B.; Lin, M.; Chen, T.; Wang, L. FG-BERT: a generalized and self-supervised functional group-based molecular representation learning framework for properties prediction. *Briefings in Bioinformatics* 2023, 24, bbad398.
- [32] Wang, X.; Wang, C.; Ji, B.; Wang, J.; Zheng, M.; Song, L.; Peng, S.; Shang, X. Multimodal pre-training models of molecular representation for drug discovery. *National Science Review* 2026, 13, nwaf495.
- [33] Nguyen, L. D.; Nguyen, Q. H.; Trinh, Q. H.; Nguyen, B. P. From SMILES to enhanced molecular property prediction: a unified multimodal framework with predicted 3D conformers and contrastive learning techniques. *Journal of Chemical Information and Modeling* 2024, 64, 9173–9195.
- [34] Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. 2019, DOI: 10.48550/arXiv.1907.11692.
- [35] Wang, S.; Guo, Y.; Wang, Y.; Sun, C.; Huang, J. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* 2019, 429–436.
- [36] Zeng, Z.; Yao, Y.; Liu, Z.; Sun, M. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature Communications* 2022, 13, 862.
- [37] Su, Q.; Ye, Z.; Han, C.; et al. QSAR prediction of BBB permeability based on machine learning upon PETBD: a novel data set of PET tracers. *Journal of Medicinal Chemistry* 2026, 69, 944–963.
- [38] Zhang, Y.; Pentikäinen, O. T. Enhancing the predictive power of macrocyclic drug permeability by knowledge distillation from analogous pretraining data. *Journal of Medicinal Chemistry* 2025, 69, 387–399.
- [39] Gupta, M.; Lee, H. J.; Barden, C. J.; Weaver, D. F. The blood–brain barrier (BBB) score. *Journal of Medicinal Chemistry* 2019, 62, 9824–9836.
- [40] Wager, T. T.; Hou, X.; Verhoest, P. R.; Villalobos, A. Moving beyond rules: the development of a central nervous system multiparameter optimization (CNS MPO) approach to enable alignment of druglike properties. *ACS Chemical Neuroscience* 2010, 1, 435–449.
- [41] Chikhale, E. G.; Ng, K. Y.; Burton, P. S.; Borchardt, R. T. Hydrogen bonding potential as a determinant of the in vitro and in situ blood–brain barrier permeability of peptides. *Pharmaceutical Research* 1994, 11, 412–419.

- [42] Banks, W. A.; Kastin, A. J. Peptides and the blood-brain barrier: lipophilicity as a predictor of permeability. *Brain Research Bulletin* 1985, 15, 287–292.
- [43] Mayhan, W. G.; Heistad, D. D. Permeability of blood-brain barrier to various sized molecules. *American Journal of Physiology-Heart and Circulatory Physiology* 1985, 248, H712–H718.
- [44] Clark, D. E. In silico prediction of blood–brain barrier permeation. *Drug Discovery Today* 2003, 8, 927–933.
- [45] Norinder, U.; Haeberlein, M. Computational approaches to the prediction of the blood–brain distribution. *Advanced Drug Delivery Reviews* 2002, 54, 291–313.
- [46] Ballet, S.; Misicka, A.; Kosson, P.; et al. Blood–brain barrier penetration by two dermorphin tetrapeptide analogues: role of lipophilicity vs structural flexibility. *Journal of Medicinal Chemistry* 2008, 51, 2571–2574.
- [47] Shityakov, S.; Neuhaus, W.; Dandekar, T. Analysing molecular polar surface descriptors to predict blood-brain barrier permeation. *International Journal of Computational Biology and Drug Design* 2013, 6, 146–156.
- [48] Xiong, B.; Wang, Y.; Chen, Y.; et al. Strategies for structural modification of small molecules to improve blood–brain barrier penetration: a recent perspective. *Journal of Medicinal Chemistry* 2021, 64, 13152–13173.
- [49] Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature Chemistry* 2012, 4, 90–98.
- [50] Xiong, Z.; Wang, D.; Liu, X.; et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry* 2019, 63, 8749–8760.
- [51] Heid, E.; Greenman, K. P.; Chung, Y.; et al. Chemprop: a machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling* 2023, 64, 9–17.
- [52] Carpenter, T. S.; Kirshner, D. A.; Lau, E. Y.; Wong, S. E.; Golbraikh, A.; Mortenson, D. E.; Carlson, H. A. A method to predict blood-brain barrier permeability of drug-like compounds using molecular dynamics simulations. *Biophysical Journal* 2014, 107, 630–641.
- [53] Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics* 2009, 1, 8.
- [54] Irwin, J. J.; Tang, K. G.; Young, J.; et al. ZINC20—a free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling* 2020, 60, 6065–6073.
- [55] Jiang, B.; Meng, L.; Zou, N.; et al. Mechanism-based pharmacokinetics-pharmacodynamics studies of harmine and harmaline on neurotransmitters regulatory effects in healthy rats: challenge on monoamine oxidase and acetylcholinesterase inhibition. *Phytomedicine* 2019, 62, 152967.
- [56] Marotta, G.; Massenzio, F.; Ortega, J. A.; et al. Exploration of the neuromodulatory properties of Fyn and GSK-3 β kinases exploiting 7-azaindole-based inhibitors. *Journal of Medicinal Chemistry* 2025, 68, 17130–17154.
- [57] Li, S.; Zhang, Y.; Deng, G.; et al. Exposure characteristics of the analogous β -carboline alkaloids harmaline and harmine based on the efflux transporter of multidrug resistance protein 2. *Frontiers in Pharmacology* 2017, 8, 541.
- [58] Kim, S.; Chen, J.; Cheng, T.; et al. PubChem 2025 update. *Nucleic Acids Research* 2025, 53, D1516–D1525.
- [59] Chandrasekhar, V.; Rajan, K.; Kanakam, S. R. S.; et al. COCONUT 2.0: a comprehensive overhaul and curation of the collection of open natural products database. *Nucleic Acids Research* 2025, 53, D634–D643.

- [60] Bjerrum, E. J. SMILES enumeration as data augmentation for neural network modeling of molecules. 2017, DOI: 10.48550/arXiv.1703.07076.
- [61] Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. 2020, DOI: 10.48550/arXiv.2010.09885.
- [62] Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is all you need. *Advances in Neural Information Processing Systems* 2017, 30.
- [63] Paszke, A.; Gross, S.; Massa, F.; et al. PyTorch: an imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* 2019, 32.
- [64] Wolf, T.; Debut, L.; Sanh, V.; et al. Transformers: state-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 2020, 38–45.
- [65] Landrum, G.; Tosco, P.; Kelley, B.; et al. rdkit/rdkit: 2024_09_1 (Q3 2024) release. Zenodo, 2024.
- [66] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 2011, 12, 2825–2830.
- [67] Venable, R. M.; Kramer, A.; Pastor, R. W. Molecular dynamics simulations of membrane permeability. *Chemical Reviews* 2019, 119, 5954–5997.
- [68] Van Der Spoel, D.; Lindahl, E.; Hess, B.; et al. GROMACS: fast, flexible, and free. *Journal of Computational Chemistry* 2005, 26, 1701–1718.
- [69] Huang, J.; Rauscher, S.; Nawrocki, G.; et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods* 2017, 14, 71–73.
- [70] Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM general force field (CGenFF) I: bond perception and atom typing. *Journal of Chemical Information and Modeling* 2012, 52, 3144–3154.
- [71] Price, D. J.; Brooks III, C. L. A modified TIP3P water potential for simulation with Ewald summation. *The Journal of Chemical Physics* 2004, 121, 10096–10103.
- [72] Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* 2008, 29, 1859–1865.
- [73] Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* 2007, 126, 014101.
- [74] Bernetti, M.; Bussi, G. Pressure control using stochastic cell rescaling. *The Journal of Chemical Physics* 2020, 153, 114107.
- [75] Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: an $N \log(N)$ method for Ewald sums in large systems. *The Journal of Chemical Physics* 1993, 98, 10089–10092.
- [76] Grubmüller, H.; Heller, H.; Windemuth, A.; Schulten, K. Generalized verlet algorithm for efficient molecular dynamics simulations with long-range interactions. *Molecular Simulation* 1991, 6, 121–142.
- [77] Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: a linear constraint solver for molecular simulations. *Journal of Computational Chemistry* 1997, 18, 1463–1472.
- [78] Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* 1993, 98, 5648–5652.
- [79] Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry* 1992, 13, 1011–1021.

- [80] Yang, J. Y.; Shin, D.-S.; Jeong, M.; Kim, S. S.; Jeong, H. N.; Lee, B. H.; Hwang, K.-S.; Son, Y.; Jeong, H.-C.; Choi, C.-H.; et al. Evaluation of drug blood-brain-barrier permeability using a microfluidic chip. *Pharmaceutics* 2024, 16, 574.

TOC Graphic



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supportinginformation.pdf](#)
- [B3DBregression.csv](#)
- [BBBPclassification.csv](#)
- [pubchemcoconut1.7M.csv](#)
- [20250115L4700CNSPenetrantCompoundLibrary96well.xlsx](#)