

Supporting Information

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.

Contents

Supplementary Figures

Figure S1	UMAP plots of pre-training chemical space after enumeration.	3
Figure S2	Visualization of exemplified natural product molecules from 7 different pathway classifiers of the COCONUT library.	4
Figure S3	Performances of model variants adapted upon BBBP classification data set.	5
Figure S4	Scatter plot of the R^2 correlation between tested BBBP labels (n=105, 10% B3DB regression data set) and predicted scores.	6
Figure S5	Distributions of physicochemical properties of top-1000 molecules across baseline models.	7
Figure S6	SynNat-BERT scoring distributions of screening molecules.	8
Figure S7	Distribution of synthetic accessibility scores and maximum similarities of CNS penetrants.	9
Figure S8	Transwell assay concentration curves and cytotoxicity results.	10
Figure S9	Snapshots of representative conformations extracted from exemplified transmembrane MD simulations of three tested alkaloids.	11
Figure S10	Visualization of attention weights of the tested candidates.	12
Figure S11	Self-correlation cluster map of bulk molecular fingerprints	22
Figure S12	Contributions of descriptor modality and individual descriptors.	23
Figure S13	SHAP analysis of feature contributions.	24
Figure S14	Visualization of the attention weights of previously reported structurally modified compounds with <i>in vivo</i> permeability measurements.	25
Figure S15	Chemical structures and prediction scores of molecules possessing experimentally tested scaffolds.	26

Supplementary Tables

Table S1	Tokenization comparison between different pre-training data set.	13
Table S2	Recovery comparison of the pre-trained SynNat-BERT	14
Table S3	Ablation studies of SynNat-BERT upon BBBP prediction accuracy.	15
Table S4	Comparison of BBBP classification performance.	16
Table S5	Comparison of BBBP regression performance.	17
Table S6	P_{app} of seven tested compounds (n=3).	18
Table S7	Hyperparameter settings of the adapted SynNat-BERT.	18
Table S8	Hyperparameter settings of FG-BERT.	19
Table S9	Hyperparameter settings of SPE-BERT.	19
Table S10	Hyperparameter settings of other baselines.	20
Table S11	Incorporated interpretable features and their sources.	21
Table S12	Predictive accuracy comparison.	22

Supplementary Notes

Note S1	Interpretable feature engineering of SynNat-BERT	21
Note S2	SynNat-BERT predicted activity cliff	24

1 Supplementary Figures

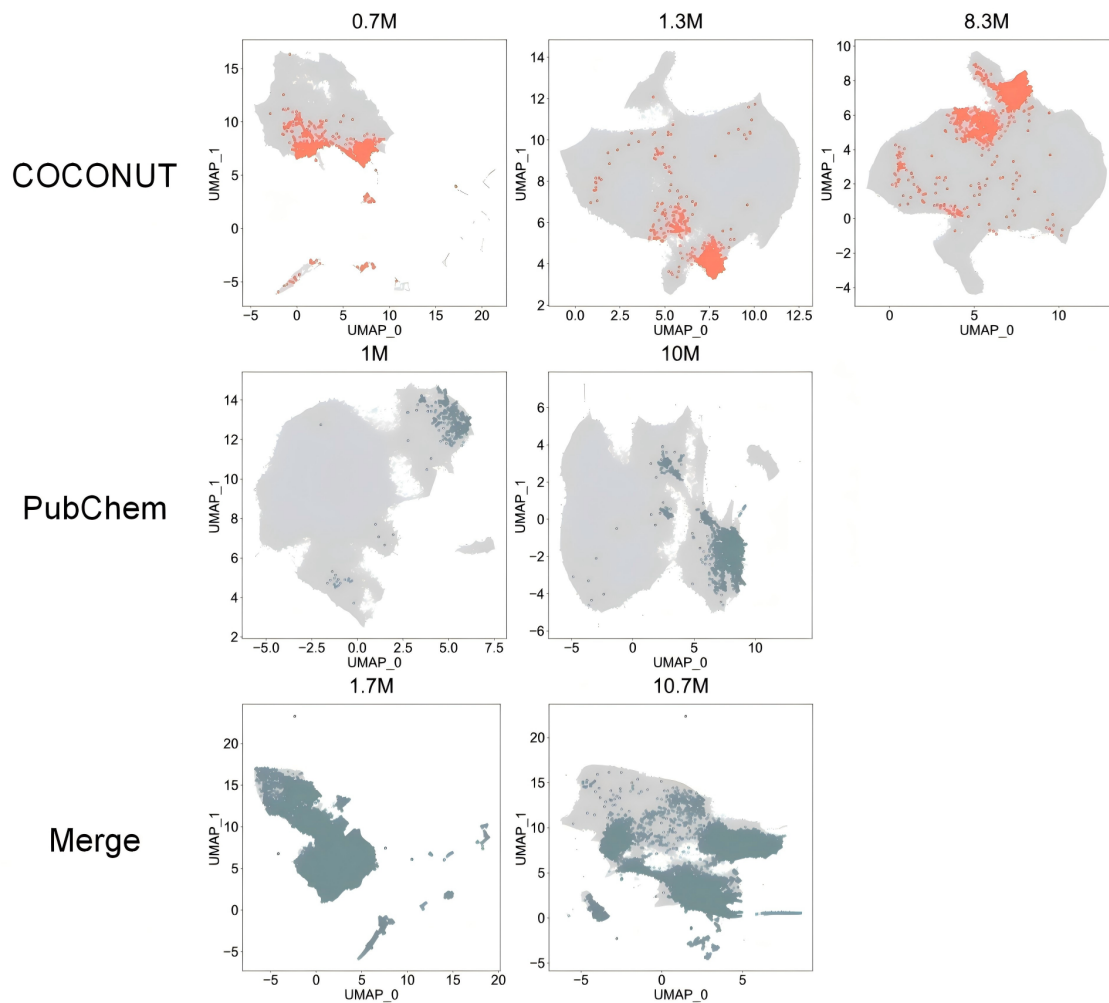


Figure S1: UMAP plots of pre-training chemical space after enumeration. All of the chemical spaces are generated by the pre-trained SynNat-BERT model. Molecules overlapped between two databases are labeled. Colors correspond to Figure 2b in the main script (red: chemical synthesis; blue: natural product). With the integration of COCONUT library, the increase of chemical space validates the enumeration strategy could enhance molecular diversity through expanding semantic representations.

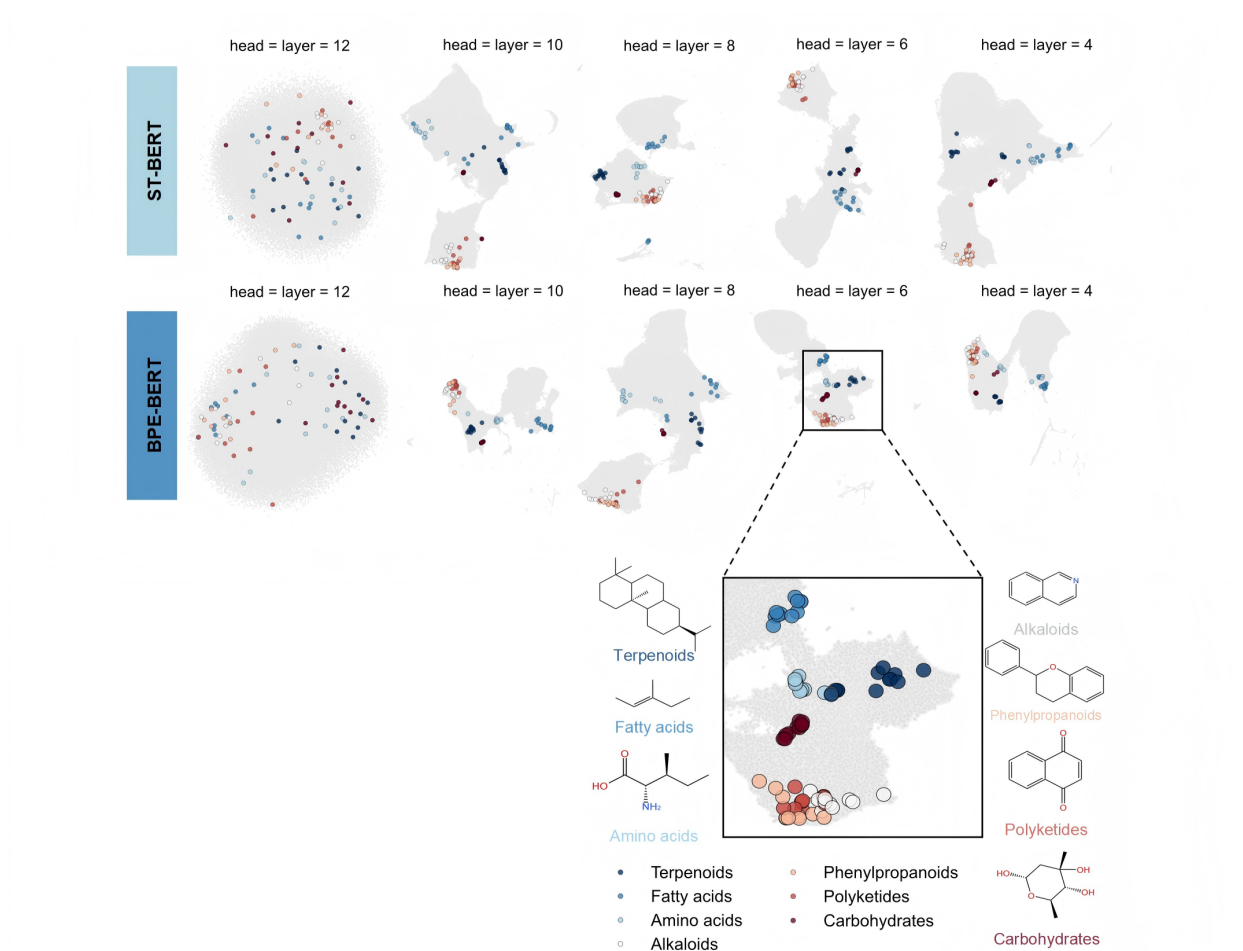


Figure S2: Visualization of exemplified natural product molecules from 7 different pathway classifiers of the COCONUT library. In the low-dimensional embedding space generated by larger models, representative natural products were uniformly distributed, suggesting a substantial risk of overfitting. Saturated cyclic molecules, predominantly occupied the middle region in the zoomed embedding space, whereas aromatic categories e.g., alkaloids, phenylpropanoids, and polyketides mainly distributed in the lower area (lower panel), indicating that the moderate model pre-trained with BPE tokenizer (head=laryer=6) possessed certain capacity to distinguish saturated ringe system from unsaturated scaffold.

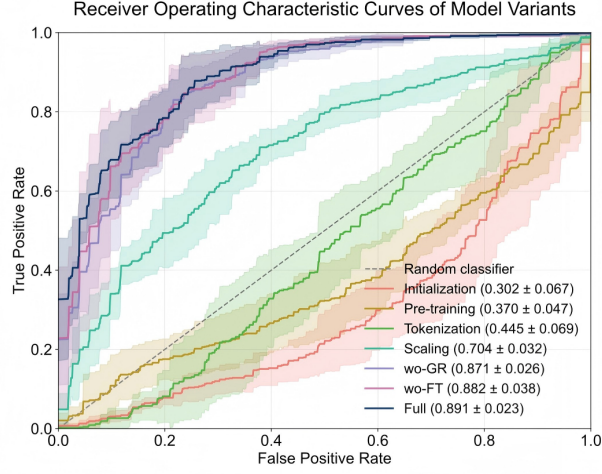
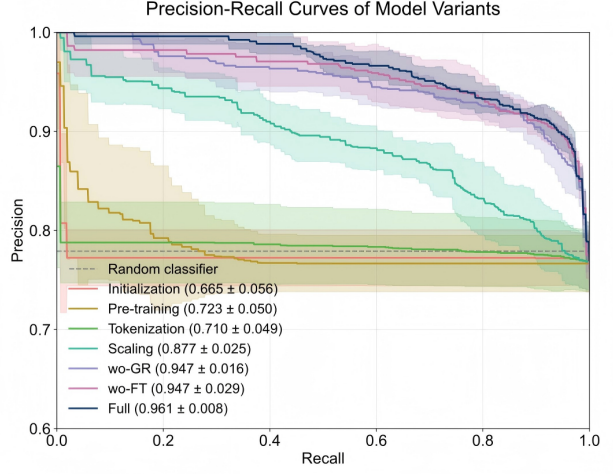
A**B**

Figure S3: Performances of model variants adapted upon BBBP classification data set. (A) Receiver operating characteristic curves (ROC) ; (B) Precision-recall curves (PRC) to evaluate the classification capabilities of different model variants. Areas under ROC and PRC are labeled to indicate AUROC and AUPRC values. Errors derive from five-fold cross-validation adaptation ($n=5$). Shaded areas are indicated by five-fold cross-validation ($n=5$).

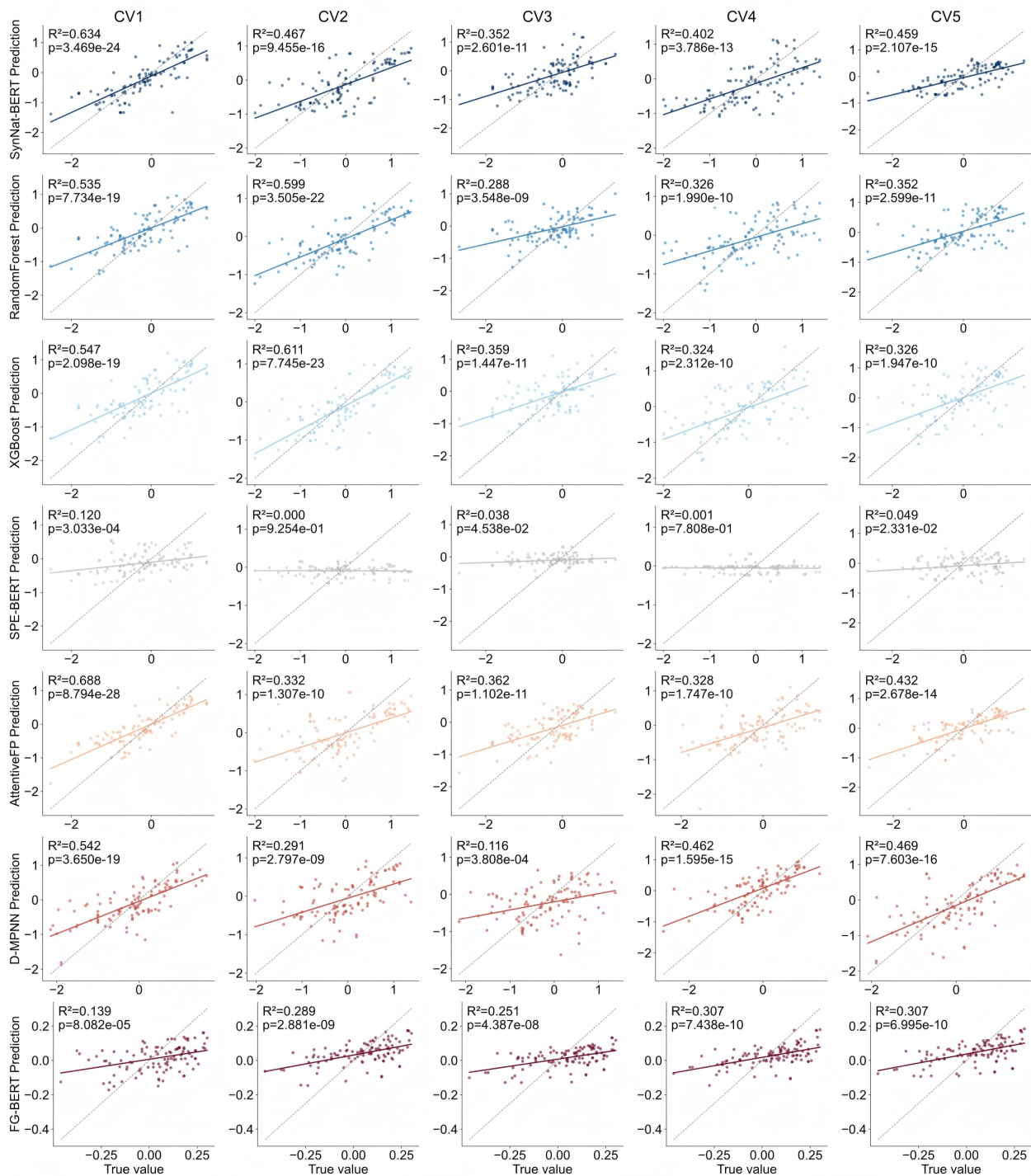


Figure S4: Scatter plot of the R^2 correlation between tested BBBP labels ($n=105$, 10% B3DB regression data set) and predicted scores. Optimal correlation is indicated by gray dashed lines. Solid line with corresponding colors denotes the linear regression fit. Colors correspond to **Fig. 4** in the main text.

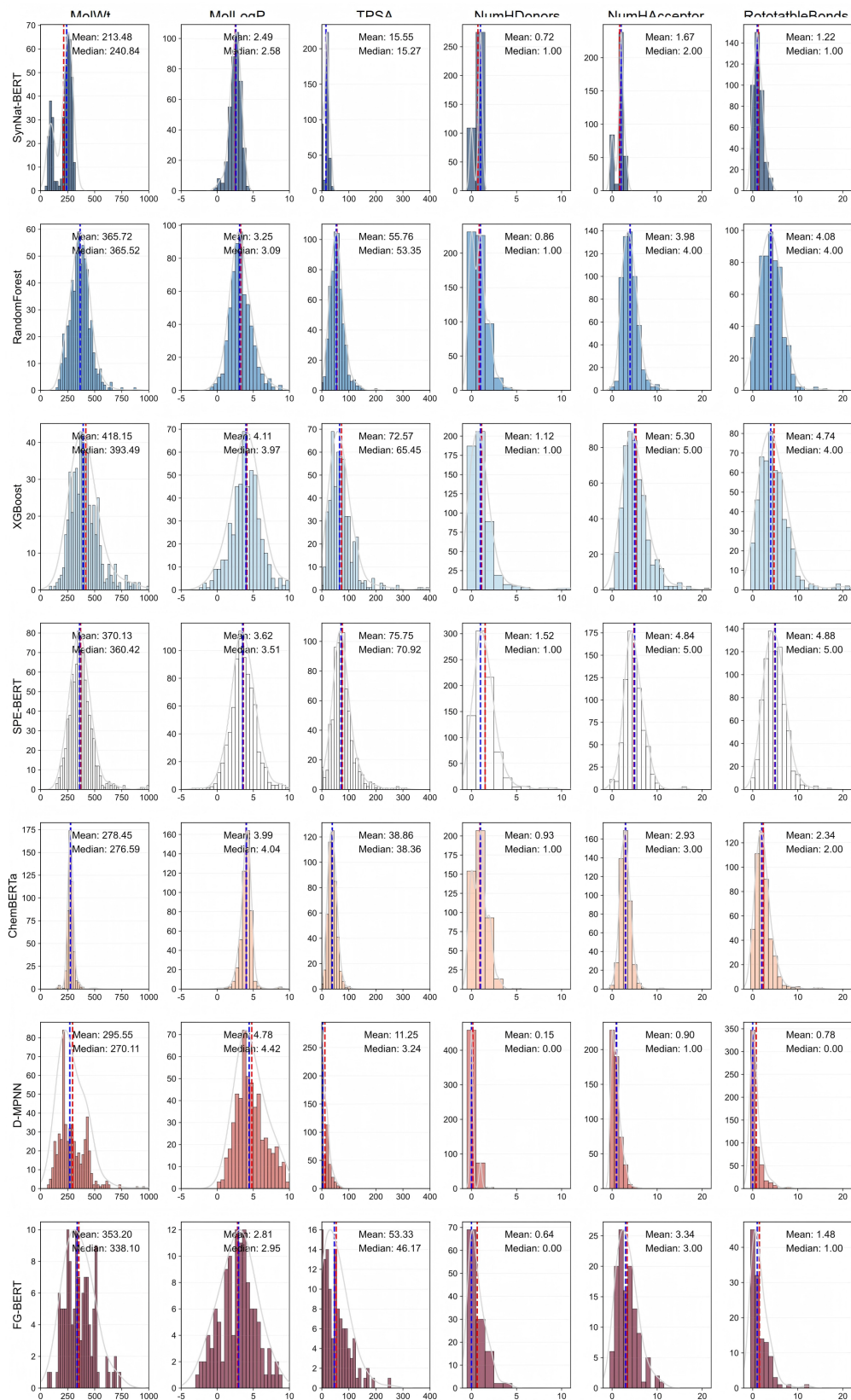


Figure S5: Distributions of physicochemical properties of top-1000 molecules across baseline models. Mean and dashed lines, respectively. Model's colors correspond to **Fig. 4** in the main text.

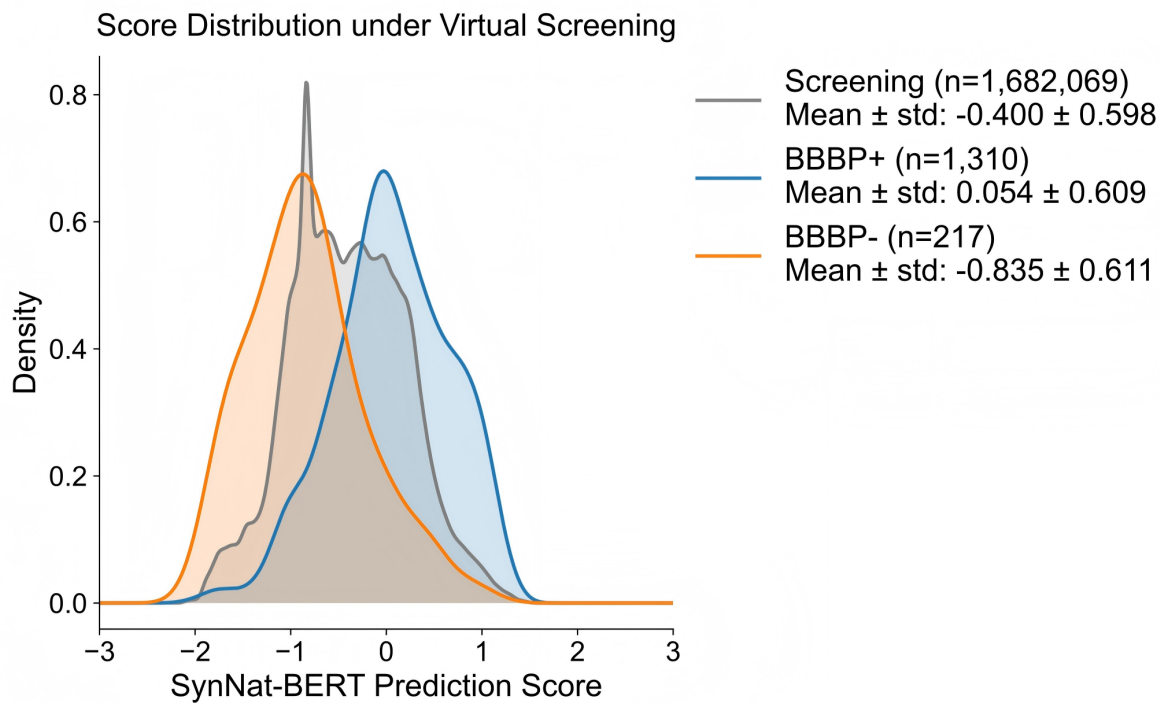


Figure S6: SynNat-BERT scoring distributions of screening molecules. Positive (blue) and negative (orange) samples from the MoleculeNet BBBP classification data set are overlapped. Screening molecules receiving positive prediction scores highly overlapped with BBB-permeable compounds suggests the rationale of selecting prediction $\log\text{BB} > 0$ as permeable threshold.

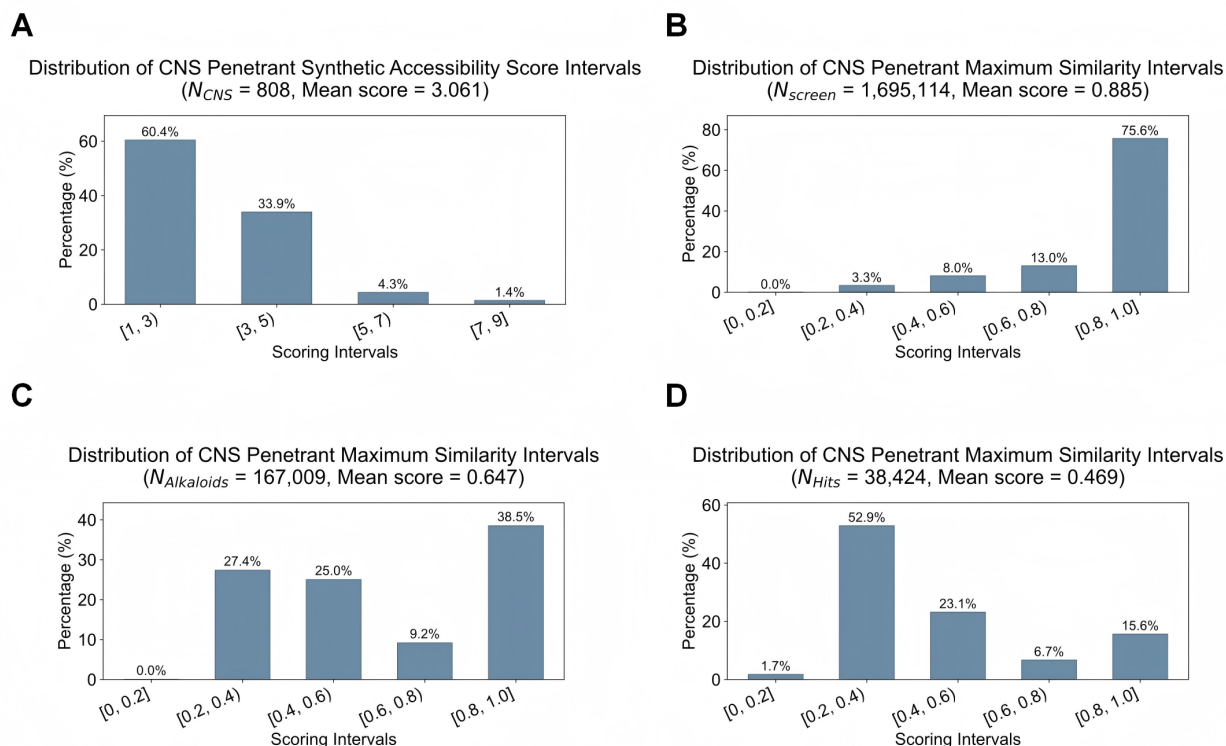
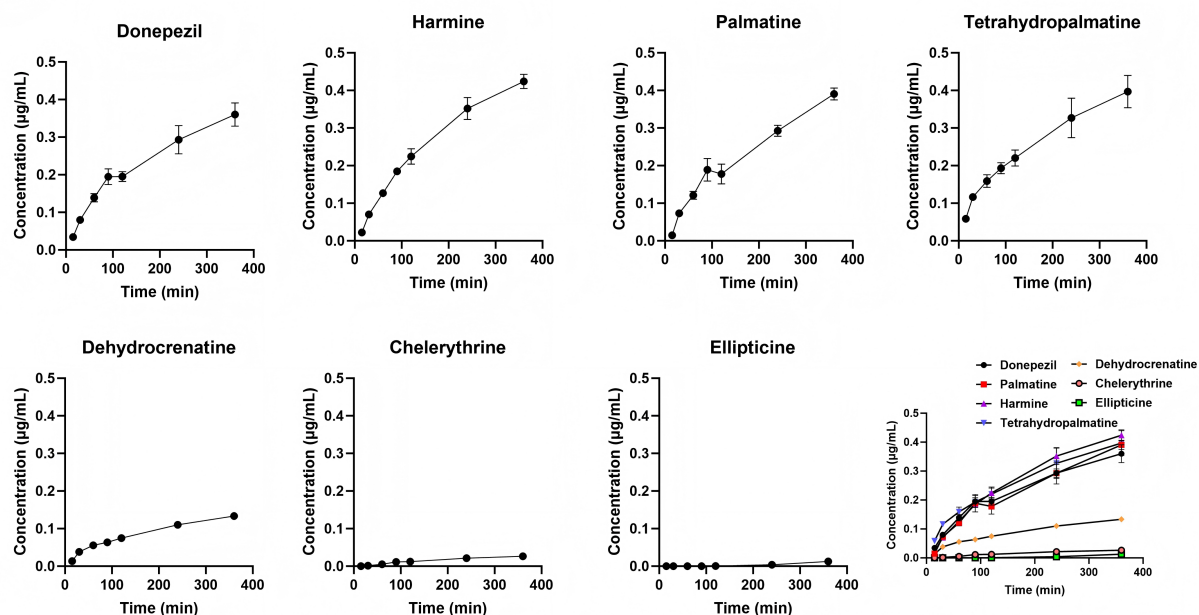


Figure S7: Distribution of synthetic accessibility scores and maximum similarities of CNS penetrants. (A) Calculated synthetic accessibility scores of CNS penetrants suggest the synthetic accessibility threshold of <3 involves over 60% potential CNS penetrable compounds. (B-D) Calculated maximum similarities of CNS penetrants towards compounds in screening library (B), alkaloid natural products (C), and potential alkaloid hits (D). These results indicate that the Tanimoto similarity threshold of < 0.4 simultaneously enables rapid screening of structurally dissimilar hits (3.3%) while maintaining structural diversity (52.9%), aligning well with the objective of our proof-of-concept screening. CNS penetrable library is derived from an external validation (SelleckChem, USA).

A



B

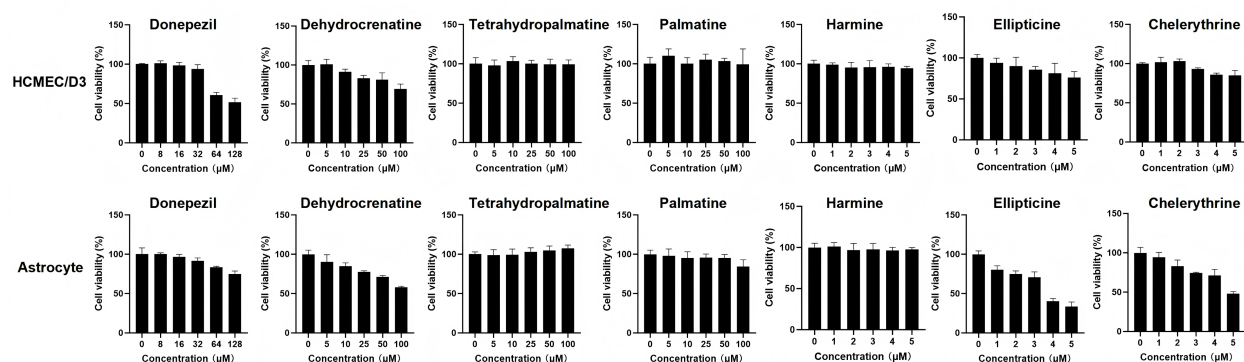


Figure S8: (A) Concentration versus dosing time of seven tested molecules on the basolateral side. (B) Cytotoxicity of seven tested molecules on HCMEC/D3 and AST cells.

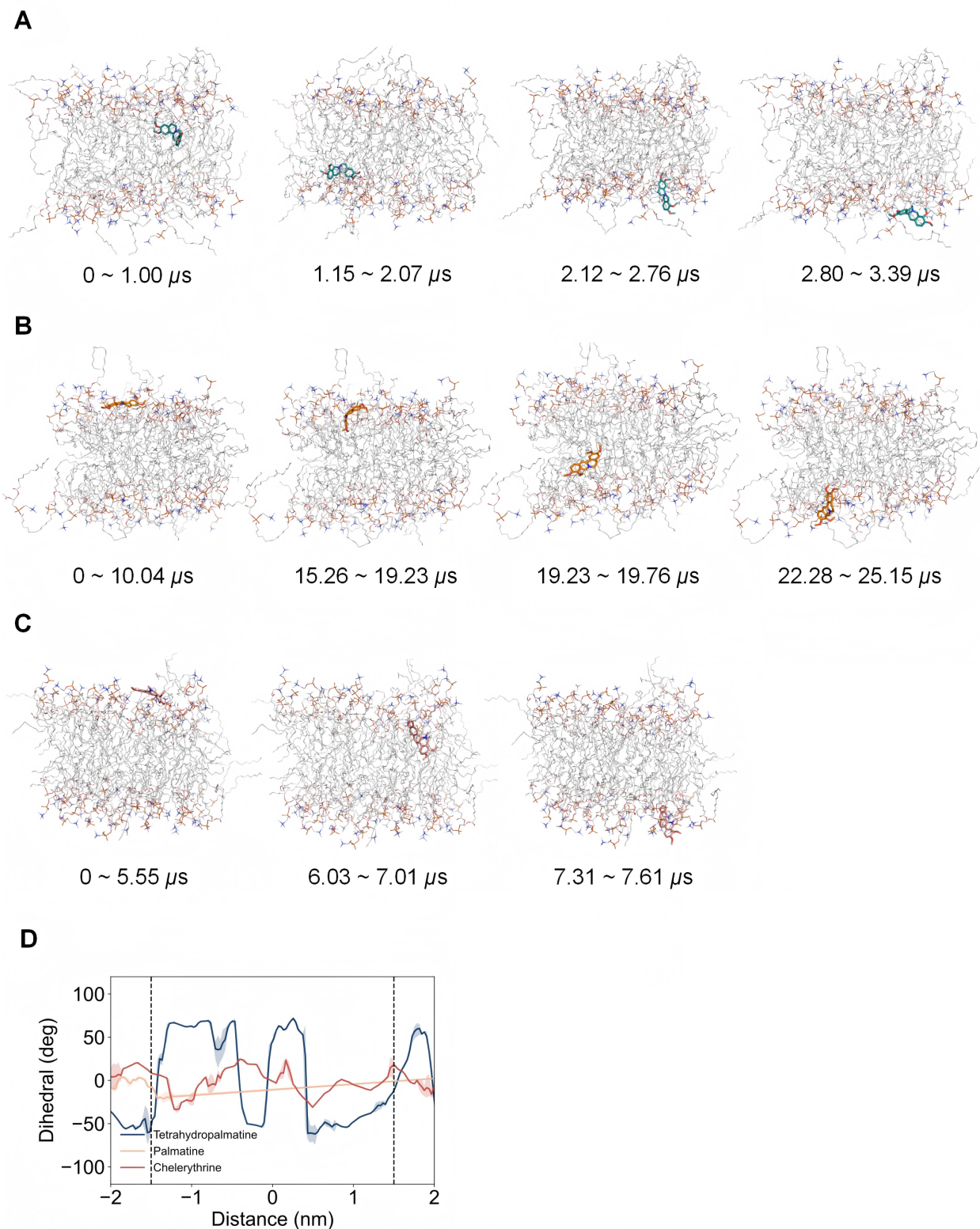


Figure S9: Snapshots of representative conformations extracted from exemplified transmembrane MD simulations of three tested alkaloids. (A) tetrahydropalmatine, (B) Palmatine, and (C) Chelerythrine. Simulation time intervals to traverse through a specific conformation are labeled during umbrella sampling simulations. Longer residence times in a conformational state suggest higher energy barriers required for conformational transition. (D) Fluctuation of dihedral angles of three tested alkaloids to reflect the overall flexibility of the scaffold. The flexible backbone of tetrahydropalmatine ($-50^{\circ} \sim 50^{\circ}$) enables it to quickly pass through bilayer cell membrane. Shaded areas are indicated by repeated sampling at a specific position. Black dashed lines in $x=-1.5$ and $x=1.5$ indicate the positions of bilayer membrane in x -axis.

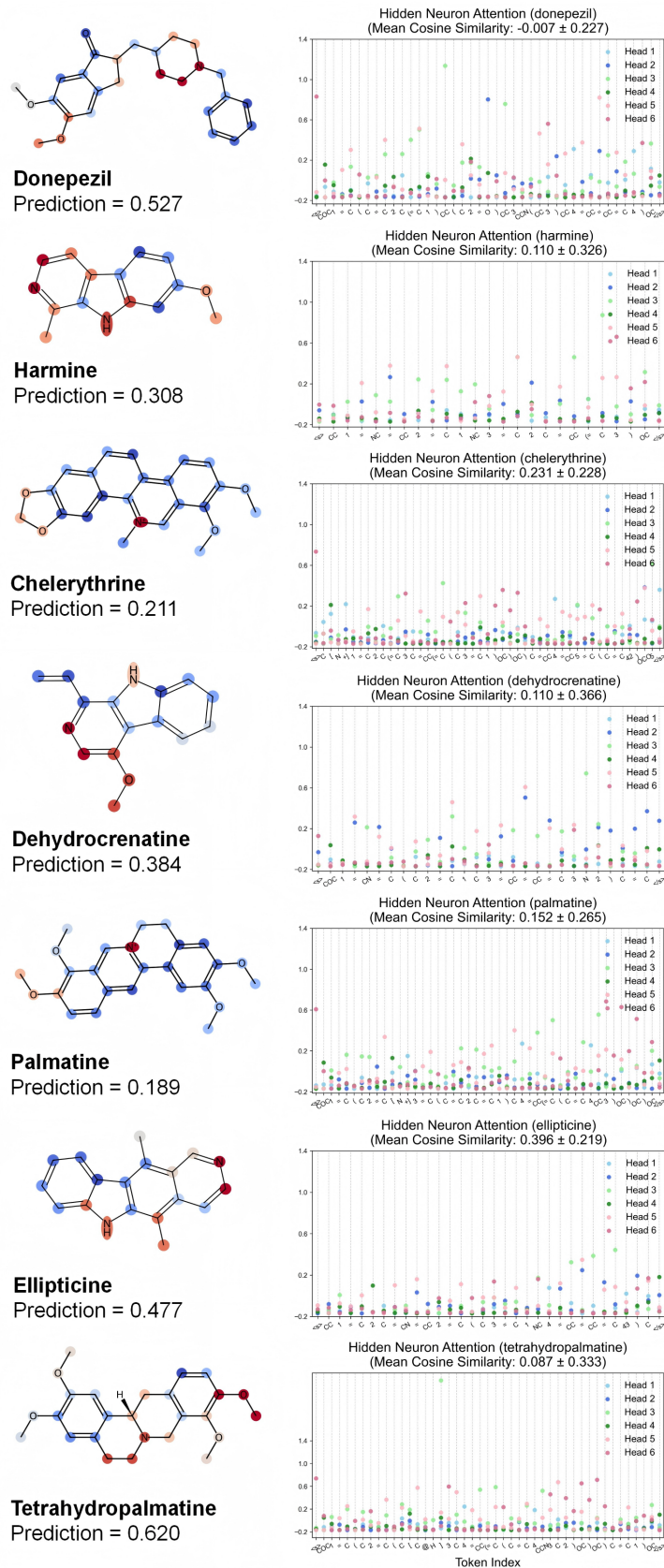


Figure S10: Visualization of attention weights of the tested candidates. Atoms colored in red represents fragments significantly contributed to BBBP property. Attention scores across varying attention heads are visualized in the right panel. Mean cosine similarity (m.c.s.) is quantified to compare the overall similarity between different attention heads.

2 Supplementary Tables

Table S1: Tokenization comparison between different pre-training data set.

Tokenization	Pre-training Data sets	Trainable Parameters (M)	Recovery Accuracy (r.a.)
Byte-Pair Encoding (BPE)	Merged 1.7 M	63.7	0.843 ± 0.012
	Merged 10.7 M	63.7	0.831 ± 0.012
	Merged 29.7 M	63.7	0.797 ± 0.019
	PubChem 1M	63.7	0.803 ± 0.012
	PubChem 10M	63.7	0.813 ± 0.018
	COCONUT 0.7 M	63.7	0.587 ± 0.023
	COCONUT 1.3M	63.7	0.674 ± 0.020
	COCONUT 8.3 M	63.7	0.690 ± 0.018
	COCONUT 19 M	63.7	0.703 ± 0.029
SMILES-tokenizer (ST)	Merged 1.7 M	60.8	0.683 ± 0.017
	Merged 10.7 M	60.8	0.834 ± 0.014
	Merged 29.7 M	60.8	0.777 ± 0.016
	PubChem 1M	60.8	0.802 ± 0.015
	PubChem 10M	60.8	0.819 ± 0.022
	COCONUT 0.7 M	60.8	0.411 ± 0.015
	COCONUT 1.3M	60.8	0.615 ± 0.016
	COCONUT 8.3 M	60.8	0.642 ± 0.011
	COCONUT 19 M	60.8	0.614 ± 0.023

* Recovery accuracy (r.a.) of SynNat-BERT with varying pre-training data sets. Merged data sets are generated from collections of PubChem + COCONUT databases. All values are presented by mean $\pm$ standard deviation (std) derived from five-fold cross-validations. Best performances for each tokenizer are **bold**.

Table S2: Recovery comparison of the pre-trained SynNat-BERT

Baselines	Trainable Parameter (M)	COCONUT (0.7 M)	Recovery Accuracy (r.a.) PubChem (1.7 M)	Zinc (250 k)	ChemBL (2.8 M)
SmilesTokenizer	ChemBERTa_zinc250k_40k[l]	83.5	0.599 $\pm$ 0.036	0.879 $\pm$ 0.014	0.523 $\pm$ 0.028
ChemBERTa-100M-MLM[2]	92.1	0.581 $\pm$ 0.038	0.701 $\pm$ 0.039	0.636 $\pm$ 0.015	0.460 $\pm$ 0.010
PubChem10M	SMILES_BPE_450k	83.5	0.414 $\pm$ 0.022	0.631 $\pm$ 0.030	0.417 $\pm$ 0.031
ChemBERTa_PubChem1M_shard00_155k	83.5	0.436 $\pm$ 0.028	0.596 $\pm$ 0.015	0.666 $\pm$ 0.029	0.428 $\pm$ 0.025
ChemBERTa-77M-MTR	3.43	0.004 $\pm$ 0.002	0.009 $\pm$ 0.007	0.003 $\pm$ 0.002	0.006 $\pm$ 0.002
ChemBERTa-ZINC-base-v1	44.1	0.009 $\pm$ 0.005	0.008 $\pm$ 0.003	0.006 $\pm$ 0.002	0.007 $\pm$ 0.003
BPE-BERT-1.7M (vocabulary size 52000)	80.6	0.849 $\pm$ 0.010	0.843 $\pm$ 0.012	0.739 $\pm$ 0.028	0.533 $\pm$ 0.021
BPE-BERT-1.7M (vocabulary size 7924)	63.7	0.844 $\pm$ 0.014	0.834 $\pm$ 0.011	0.736 $\pm$ 0.018	0.522 $\pm$ 0.016
BPE-BERT-10.7M	80.6	0.620 $\pm$ 0.023	0.779 $\pm$ 0.019	0.447 $\pm$ 0.013	0.342 $\pm$ 0.014
BPE-BERT-29.7M	80.6	0.683 $\pm$ 0.015	0.719 $\pm$ 0.018	0.620 $\pm$ 0.032	0.441 $\pm$ 0.030

* Recovery accuracy (r.a.) of the pre-trained Robustly optimized BERT approach (RoBERTa, above) with our pre-trained SynNat-BERT variants (below). Number of molecular samples used to measure r.a. is represented in bracket. All values are presented as mean $\pm$ std derived from five-fold cross-validations. Best performances of each data set are **bold**.

Table S3: Ablation studies of SynNat-BERT upon BBBP prediction accuracy.

Variants	Trainable Parameters (M)	Self Pretrained	BBBP Adapted	AUROC	AUPRC	MCC	Sensitivity	F1
ChemBERTa-100M-MLM	92.1	×	×	0.302 ± 0.067	0.665 ± 0.056	0.064 ± 0.079	0.391 ± 0.053	0.577 ± 0.479
Pre-training (10.7 M; ST)	60.8	✓	×	0.370 ± 0.047	0.723 ± 0.050	0.073 ± 0.060	0.102 ± 0.082	0.174 ± 0.134
Tokenization (10.7 M; BPE)	63.7	✓	×	0.445 ± 0.069	0.710 ± 0.049	0.061 ± 0.056	0.615 ± 0.385	0.741 ± 0.090
Scaling (1.7 M; BPE)	63.7	✓	×	0.704 ± 0.032	0.877 ± 0.025	0.341 ± 0.044	0.682 ± 0.100	0.764 ± 0.060
w/o GR	124.6	✓	✓	0.871 ± 0.026	0.948 ± 0.016	0.562 ± 0.064	0.924 ± 0.035	0.902 ± 0.013
w/o FT	124.6	✓	✓	0.882 ± 0.038	0.948 ± 0.029	0.628 ± 0.082	0.944 ± 0.026	0.910 ± 0.011
Full	124.6	✓	✓	0.891 ± 0.023	0.961 ± 0.008	0.590 ± 0.098	0.938 ± 0.035	0.913 ± 0.015

* Classification performance of SynNat-BERT compared to its variants. The above models are evaluated under inference mode. Below are adaptation models based on BBBP classification data set, pre-trained on 1.7 M data set using BPE tokenizer. Descriptors are incorporated through standardized cross-attention fusion (w/o Gaussian reweighted), denoted as w/o GR), Gaussian reweighted cross-attention fusion with text encoder (w/o freeze text, denoted as w/o FT), Gaussian reweighted cross-attention fusion with freezing text (denoted as full). All values are presented as mean ± std derived from five-fold cross-validations. Best performances for each metrics are **bold**.

Table S4: Comparison of BBBP classification performance.

Data sets	Metrics	Baseline Models					D-MPNN	AttentiveFP	XGBoost	Random Forest
		SynNat-BERT	SPE-BERT	FG-BERT						
MoleculeNet n=2,042	AUROC	0.891 ± 0.026	0.780 ± 0.039	0.660 ± 0.017	0.772 ± 0.030	0.815 ± 0.033	0.878 ± 0.018	0.880 ± 0.017		
	AUPRC	0.961 ± 0.008	0.923 ± 0.016	0.691 ± 0.018	0.902 ± 0.029	0.921 ± 0.020	0.940 ± 0.025	0.948 ± 0.018		
	MCC	0.590 ± 0.098	0.265 ± 0.056	0.158 ± 0.024	0.456 ± 0.078	0.569 ± 0.059	0.596 ± 0.071	0.533 ± 0.039		
	Sensitivity	0.938 ± 0.035	0.926 ± 0.036	0.804 ± 0.033	0.914 ± 0.071	0.964 ± 0.020	0.926 ± 0.017	0.959 ± 0.013		
	F1	0.913 ± 0.015	0.862 ± 0.020	0.671 ± 0.009	0.879 ± 0.033	0.913 ± 0.008	0.910 ± 0.009	0.904 ± 0.013		

* Classification performance of SynNat-BERT compared to baseline models. Adaptations are performed under 8:1:1 train-test-validation split based on Bemis-Murcko scaffold. All values are presented as mean $\pm$ std derived from five-fold cross-validations. Best performances for each metrics are **bold**.

Table S5: Comparison of BBBP regression performance.

Data set	Metrics	Baseline Models							
		SynNat-BERT	SPE-BERT	FG-BERT	D-MPNN	AttentiveFP	XGBoost	Random Forest	
B3DB n=1,058	PCC	0.677 ± 0.070	0.148 ± 0.143	0.504 ± 0.076	0.596 ± 0.161	0.647 ± 0.107	0.652 ± 0.101	0.641 ± 0.105	
	Spearman	0.702 ± 0.084	0.181 ± 0.136	0.492 ± 0.058	0.632 ± 0.138	0.662 ± 0.112	0.653 ± 0.094	0.636 ± 0.091	
	RMSE ↓	0.610 ± 0.042	0.787 ± 0.043	0.149 ± 0.007	0.626 ± 0.058	0.607 ± 0.065	0.614 ± 0.091	0.616 ± 0.073	
	MAE ↓	0.476 ± 0.042	0.633 ± 0.048	0.121 ± 0.007	0.478 ± 0.046	0.474 ± 0.057	0.451 ± 0.048	0.461 ± 0.039	
	R2	0.414 ± 0.083	0.006 ± 0.057	0.218 ± 0.073	0.334 ± 0.199	0.413 ± 0.131	0.402 ± 0.151	0.399 ± 0.137	

* Regression performance of SynNat-BERT compared to baseline models. Adaptations are performed under 8:1:1 train-test-validation split based on Bemis-Murcko scaffold. All values are presented as mean $\pm$ std derived from five-fold cross-validations. Best performances for each metrics are **bold**.

Table S6: P_{app} of seven tested compounds (n=3).

Compounds	Mean (cm/s)	SD (cm/s)
Harmane (natural product positive control)	2.26593e-06	4.77486e-07
Donepezil (chemical synthetic positive control)	3.36138e-06	1.0706e-06
Tetrahydropalmatine	3.58223e-06	4.7647e-07
Palmatine	2.01048e-06	3.93774e-08
Dehydrocrenatine	2.66304e-06	4.03754e-07
Chelerythrine	1.76443e-07	1.0185e-07
Ellipticine	5.52232e-09	3.1963e-09

* Two positive controls are labeled. Compounds are listed in descending order of P_{app} values. Among five tested hits, tetrahydropalmatine, palmatine and dehydrocrenatine, demonstrating permeability over 1e-06, are recognized as potential penetrants. Best performances for each metrics are **bold**.

Table S7: Hyperparameter settings of the adapted SynNat-BERT.

Category	Hyperparameter	Hyperparameter sets	Default value
Input	Fraction of training	0.95	0.95
	Enumerate folds	{0, 2, 10, 50}	0
Tokenizer	Type	{BPE, ST}	BPE
	Max length	{128, 256, 512}	512
	Vocabulary size	{600, 7924, 52000}	52,000
Self-attention Module	Attention dropout	{0.1, 0.2, 0.3}	0.1
	Attention heads	{2, 4, 6, 8, 10, 12}	6
	Hidden size / head	{64, 256, 384}	64
SMILES Encoder	Hidden layers	{2, 4, 6, 8, 10, 12}	6
	Intermediate size	3072	3072
	Dropout	0.1	0.1
Model Pre-training	Early stopping patience	{3, 5, 10}	10
	Evaluation steps	50	50
	Learning rate	5e-5	5e-5
	Batch size	{32, 64, 128, 256, 512}	64
	MLM probability*	0.15	0.15
	Max training epochs	{3, 10, 50, 100}	100
Descriptor Preprocess	Scalar	{Standard, Gaussian}	Gaussian
	Descriptor dimension	{3, 6, 9, 50, 100, 210}	9
Cross-attention Fusion	Type	{Cross-attention, Concatenation}	Cross-attention
	Fusion hidden dimension	{128, 256, 512}	512
	Attention dropout	0.2	0.2
Text Encoder	Freeze text	{True, False}	True
	Text projection dimension	{128, 256}	256
Model Adaptation	Early stopping patience	{10, 30}	30
	Batch size	{32, 64, 128, 256, 512}	128
	Learning rate	2e-5	2e-5
	Max training epochs	{50, 100, 200}	200

* MLM probability represents the probability of masking tokens within a SMILES sequence.

Table S8: Hyperparameter settings of FG-BERT.

Category	Hyperparameter	Hyperparameter sets	Default value
Tokenizer	Self-pretrained	{True, False}	False
	Load weight*	/	bert_weightsMedium_20
Model Architecture	Layers	6	6
	Hidden dimension	256	256
	Vocabulary size	18	18
	Heads	{4, 5, 6, 7, 8}	4
	Dense dropout	[0, 0.5]	0.2
Model Training	Learning rate	[3e-5, 15e-5]	9.09e-5
	Batch size	{16, 32, 48, 64}	32
	Training epochs	20	20

* Loaded model weight can be accessed via: <https://github.com/idrugLab/FG-BERT>.

Table S9: Hyperparameter settings of SPE-BERT.

Category	Hyperparameter	Hyperparameter sets	Default value
Tokenizer	Training SPE datasets	{0.7M, 1.7M}	1.7M
	Augmentation times	1	1
	Min frequency	2,000	2,000
	Number of symbols	30,000	30,000
	Max length	128	128
Model Architecture	Embedding dimension	300	300
	Hidden layers	{2, 4, 6, 8}	4
	Heads	{2, 4, 6, 8}	4
	Dropout	{0., 0.1, 0.2}	0.1
	Feed-forward	1,200	1,200
Model Training	Learning rate	{1e-4, 1e-5}	1e-4
	Batch size	{16, 32, 48, 64}	32
	Training epochs	30	30
	Early stopping patience	5	5

Table S10: Hyperparameter settings of other baselines.

Models	Hyperparameter	Default value
AttentiveFP	Graph attention layers	2
	Num timesteps	2
	Graph feature size	200
	Dropout rate	0.16
	Max length	128
	Learning rate	0.000016
	Weight decay	0.000749
	Node feature size	39
	Edge feature size	10
D-MPNN	Batch size	50
	Depth	3
	Activation	ReLU
	Aggregation	Mean
	Epochs	5
	Number of FFN layers	2
	FFN hidden size	300
Random Forest	Max depth	20
	Number of estimators	100
XGBoost	Objective	Squared error
	Number of jobs	16

3 Supplementary Note 1. Interpretable feature engineering of SynNat-BERT

To systematically investigate the contribution of interpretable physicochemical descriptors, and determine the minimum combination of features to achieve optimal performance of BBBP prediction, we conducted a thorough feature engineering experiments.

3.1 SynNat-BERT predictive accuracies with bulk and interpretable fingerprints

As incorporating RDKit physicochemical features was a common strategy for improving the overall performance of molecular property predictions, we first integrated a set of 210 2D RDKit descriptors as a comparison and evaluated its classification performance. We initially calculated the correlation matrix between the feature importance of this set of descriptors as measured by gradient importance (g.i.). g.i. was calculated by first-order partial derivative of each model component (e.g. molecular descriptor) to assess the contribution of a specific component to model output based on gradient descending method:

$$g.i. = \left| \frac{\partial \sum_{i=1}^N y_i}{\partial \mathbf{d}'} \right|, \quad (1)$$

where y_i is the logit of the binary classification output, and $\mathbf{d}'$ is the descriptor vector ($\in \mathbb{R}^{N \times D_d}$).

Table S11: Incorporated interpretable features and their sources.

Name	Description	Suggested Range*	Reference
NumHDonors	Number of hydrogen bond donors	< 3	[3]
NumHAcceptors	Number of hydrogen bond acceptors	/	[3]
MolLogP	Octanol-water partition coefficient	2 ~ 4	[4]
MolWt	Molecular weight	< 450	[5]
NOCount	Count of (N +O) atoms	< 5	[6, 7]
NumRotatableBonds	Number of rotatable bonds	/	[8]
TPSA	Topological polar surface area	< 90 Å	[9]

* Suggested ranges indicate the physicochemical properties that a BBB penetrant compounds would possess, derived from previous work[1, 7]. “/” represents not reported, which is defined as “0~5” in this study.

As shown in Figure S11, around 50% descriptors demonstrated pearson $r > |0.5|$, validating the existing of redundant features. Subsequently, we developed an additional model incorporating 7 well-established features in previous reports (Table S11), where we found improved performance in most of the metrics over model incorporating bulk features except for sensitivity (Table S12).

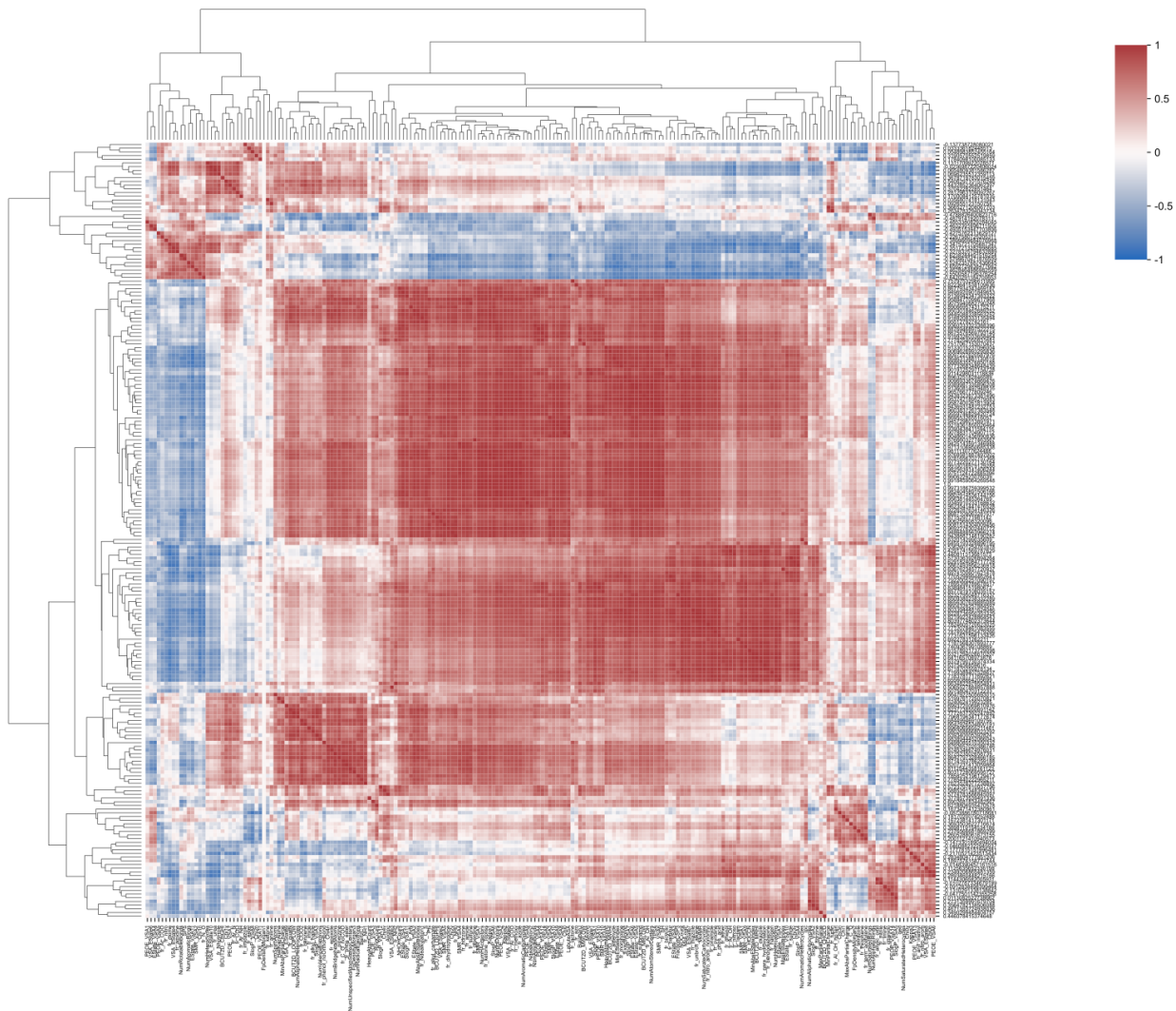


Figure S11: Self-correlation cluster map of bulk molecular fingerprints ($N_{feat} = 210$). The intensity of the color is proportional to the strength of the correlation.

Table S12: Predictive accuracy comparison.

Strategy	Num of Descriptors	AUROC	AUPRC	MCC	Sensitivity	F1
Interpretable	7	0.891 ± 0.023	0.961 ± 0.008	0.590 ± 0.098	0.938 ± 0.035	0.913 ± 0.015
Bulk	210	0.845 ± 0.063	0.933 ± 0.044	0.512 ± 0.126	0.958 ± 0.024	0.901 ± 0.021

* All values are presented as mean $\pm$ std derived from five-fold cross-validations. Best performances of each metrics are **bold**.

3.2 Feature contributions of the interpretable SynNat-BERT model

To quantitatively assess the contributions of descriptor modality and individual descriptors, we analyzed the gradient ratio associated with each model component. Compared to other model components, descriptor layer consistently took up a ratio of around 5% across all model variants. Among this layer, molecular weight (MolWt) was the most vital features across all three importance metrics (Figure S12). By using Shapley values, we further analyzed the direction and intensity of the embedded features on the output of model

incorporating all components [10]. 5-fold cross validation consistently demonstrated MolWt as the most important feature, with higher molecular weight being associated with lower model output (representing lower predicted permeability) (Figure S13). This conclusion corresponded to all three importance metrics and prior knowledge as well (Table S11). The above results validated that our developed dual-modal SynNat-BERT framework enabled quantitative attribution of property-level contributions via cross-attention mechanism encoding a descriptor layer during model adaptation process.

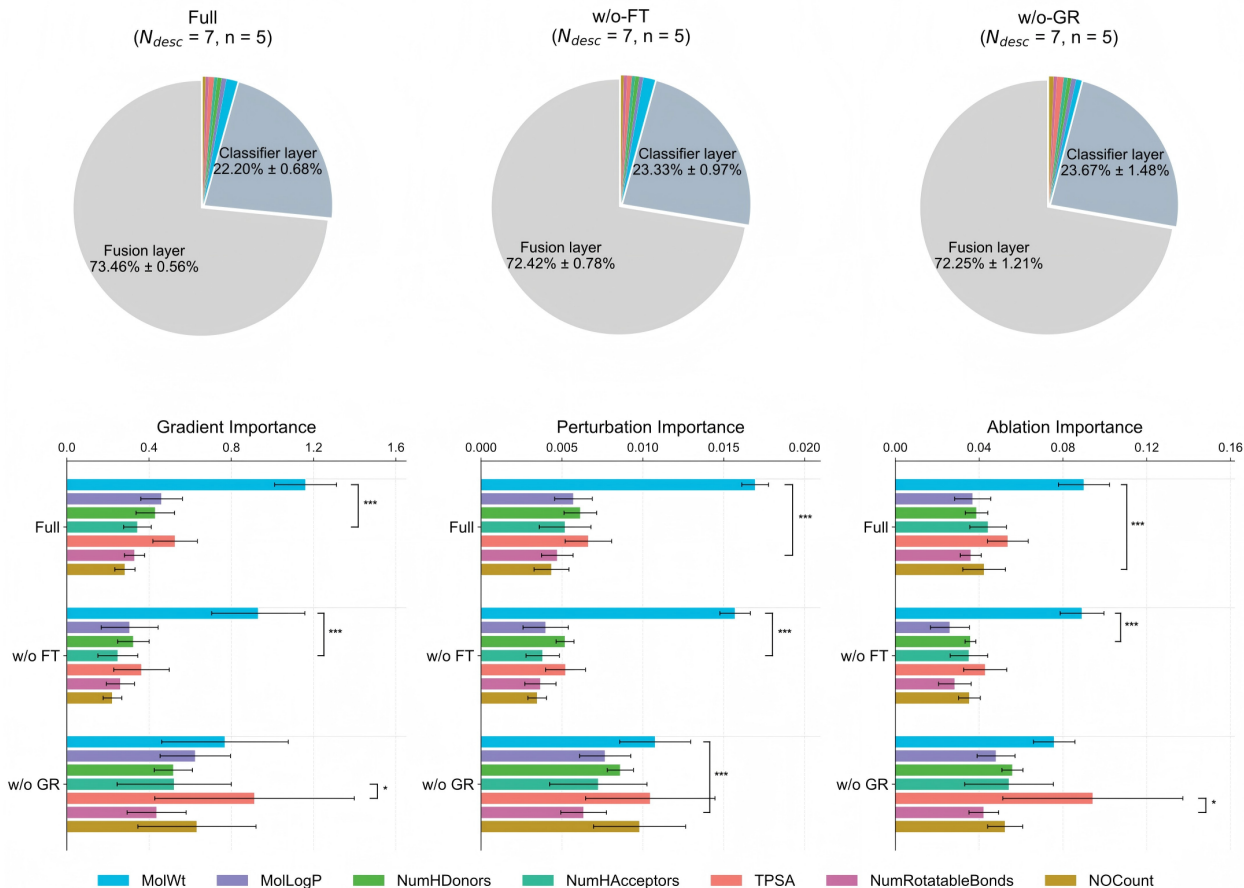


Figure S12: Contribution ratios of each model component are shown in the upper panel, while feature importance of individual descriptors are displayed in the lower panel. All values are presented as mean $\pm$ std derived from five-fold cross-validations.

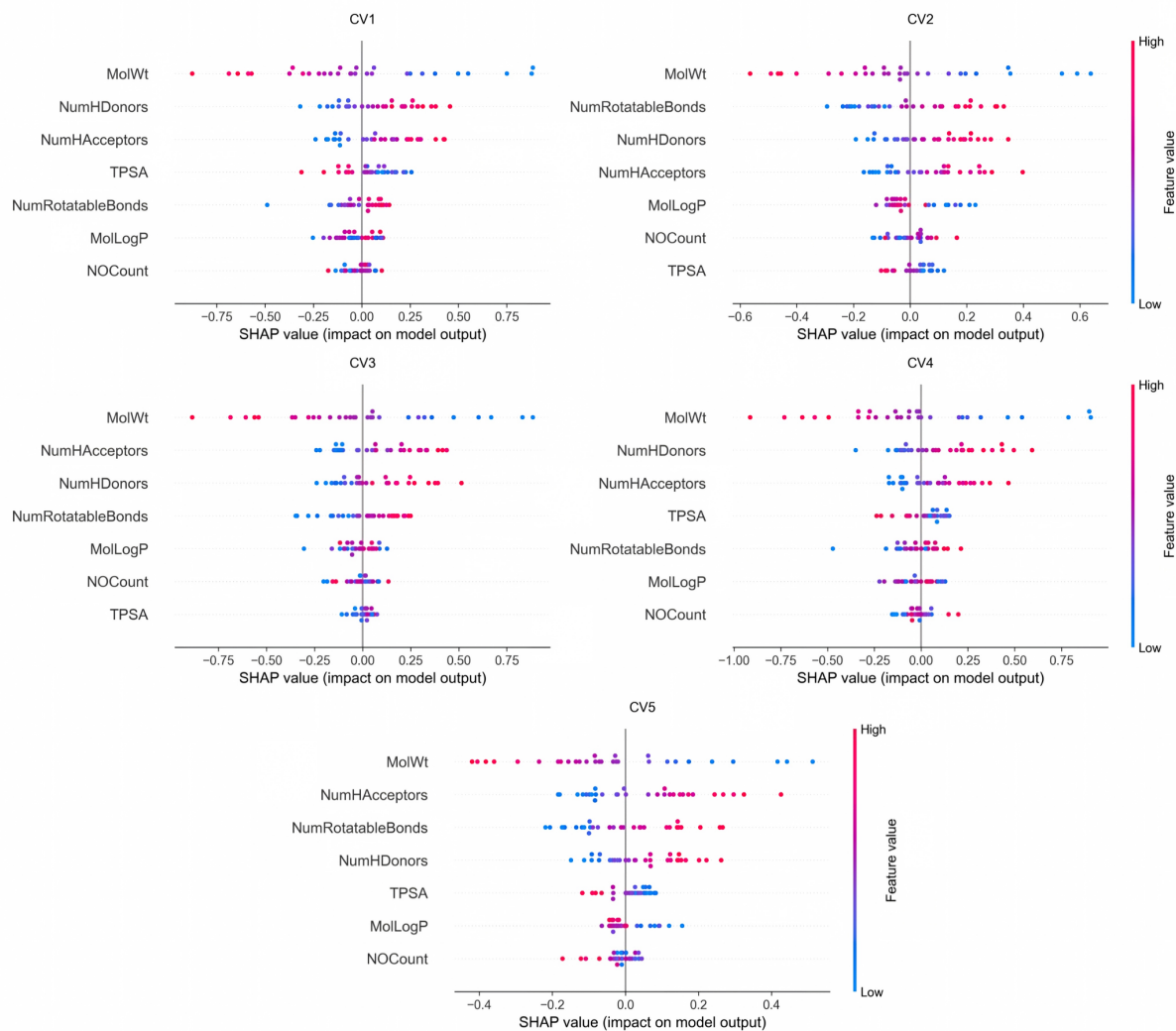


Figure S13: 5-fold cross-validation models incorporating all components (full) were used to analyze the contribution direction and intensity of each descriptor feature.

4 Supplementary Note 2. SynNat-BERT predicted activity cliff

Accurate prediction of activity cliffs requires the model to capture substantial activity differences arising from subtle structural perturbations, which are particularly challenging in predicting scaffolds outside the training set [11]. To investigate whether SynNat-BERT can distinguish bioactivity differences led by subtle structural perturbations, we compared model’s prediction scores of molecules reported in [12] and molecules possessing berbine-, protoberberine-, and benzo[c]phenanthridine-analog scaffolds within the screened candidates.

4.1 Atomic attention maps and prediction scores of unseen molecules

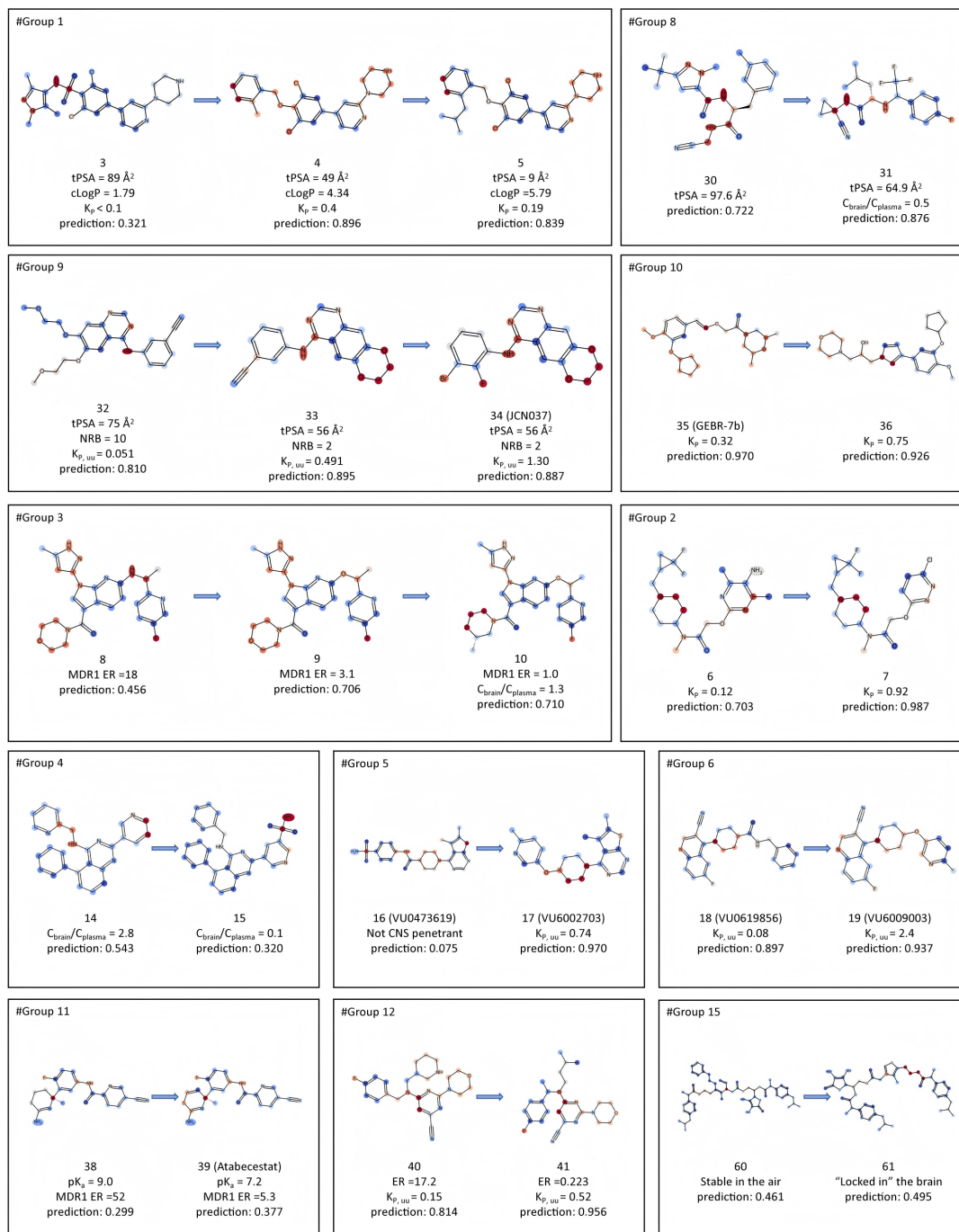


Figure S14: Visualization of the attention weights of previously reported structurally modified compounds with *in vivo* permeability measurements. *In vivo* values are derived from [12].

4.2 Activity cliff of molecules possessing benzo[c]phenanthridine, berberine, and berbine scaffolds within screened molecules

To specifically focus on the influence of scaffold, we retrieved alkaloid molecules sharing identical Bemis-Murcko scaffolds with the in-house experimentally tested compounds and excluded structures containing exposed hydroxyl groups.

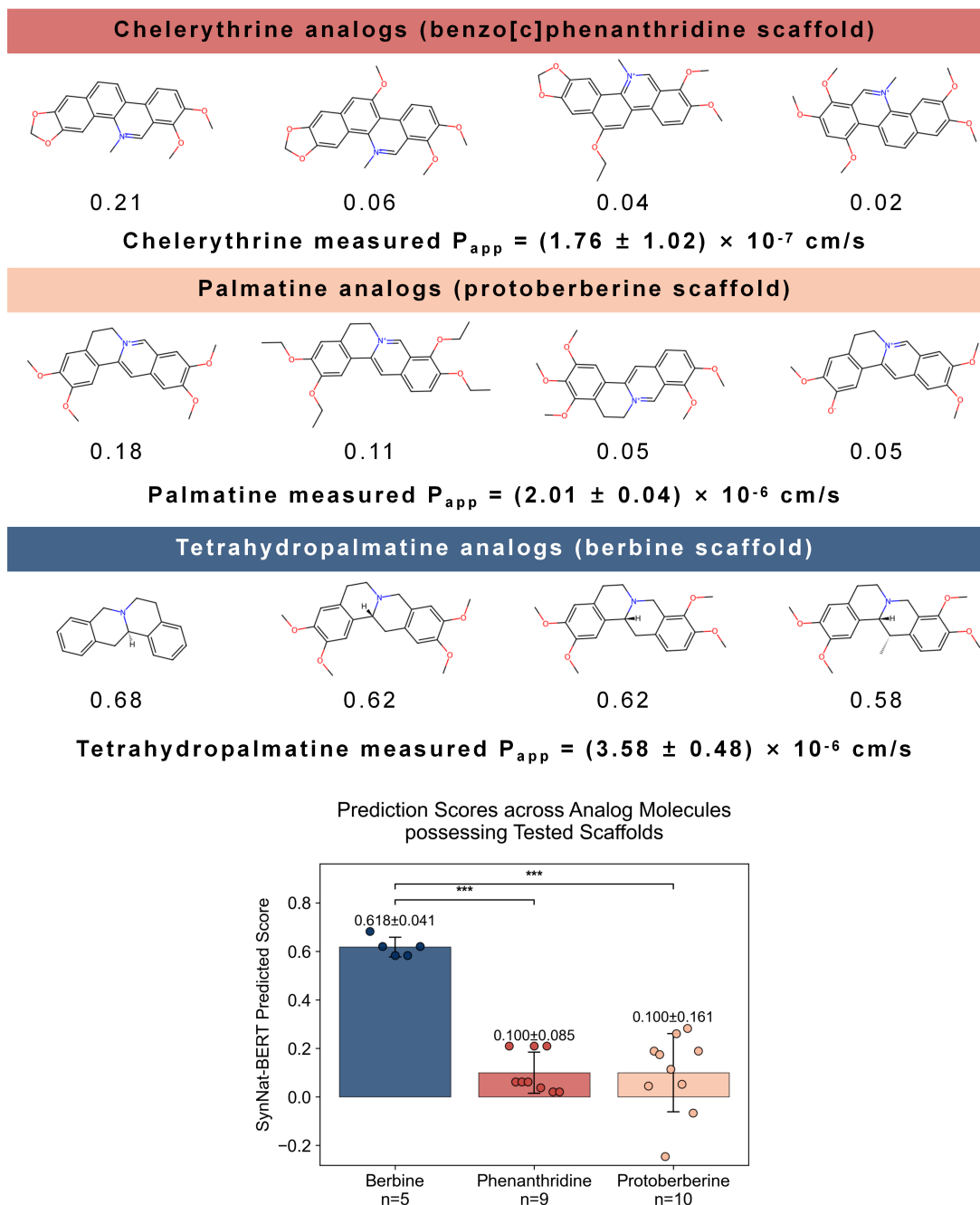


Figure S15: Examples of activity cliff chemical structures predicted by SynNat-BERT. Molecules with identical scaffolds are grouped together. The predicted logBB of each molecule is assigned. Numbers of the analog molecules are labeled below the scaffold name in the distribution plot of SynNat-BERT predicted scores.

Results demonstrated that SynNat-BERT model assigned highly consistent scores to molecules sharing the identical substructure, while producing distinct scores for molecules with different scaffolds (Figure S15). Molecules containing berbine scaffolds received relatively higher prediction scores (0.618 ± 0.041 , $n=5$), whereas aromatic protoberberine and benzo[c]phenanthridine alkaloids generally showed lower scores (0.100 ± 0.161 , $n=10$ and 0.100 ± 0.085 , $n=9$, respectively).

5 Reference

1. Chithrananda S, Grand G, Ramsundar B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885, 2020.
2. Singh R, Barsainyan A A, Irfan R, et al. ChemBERTa-3: an open source training framework for chemical foundation models. Digital Discovery **5**(2), 662-685 (2026).
3. 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 **11**(3), 412-419 (1994).
4. Banks, W.A., Kastin, A.J. Peptides and the blood-brain barrier: lipophilicity as a predictor of permeability. Brain Research Bulletin **15**(3), 287-292 (1985).
5. Mayhan, W.G., Heistad, D.D. Permeability of blood-brain barrier to various sized molecules. American Journal of Physiology-Heart and Circulatory Physiology **248**(5), 712-718 (1985).
6. Clark, D.E. In silico prediction of blood-brain barrier permeation. Drug Discovery Today **8**(20), 927-933 (2003).
7. Norinder, U., Haeberlein, M. Computational approaches to the prediction of the blood-brain distribution. Advanced Drug Delivery Reviews **54**(3), 291-313 (2002).
8. 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 **51**(8), 2571-2574 (2008).
9. 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 **6**(1-2), 146-156 (2013).
10. Winter E. The shapley value[J]. Handbook of game theory with economic applications, **3**: 2025-2054 (2003).
11. Scalia G, Rutherford S T, Lu Z, et al. Deep-learning-based virtual screening of antibacterial compounds. Nature biotechnology, 1-14 (2025).
12. 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 **64**(18), 13152-13173 (2021).