Supplementary Materials

1 PEMAL Details

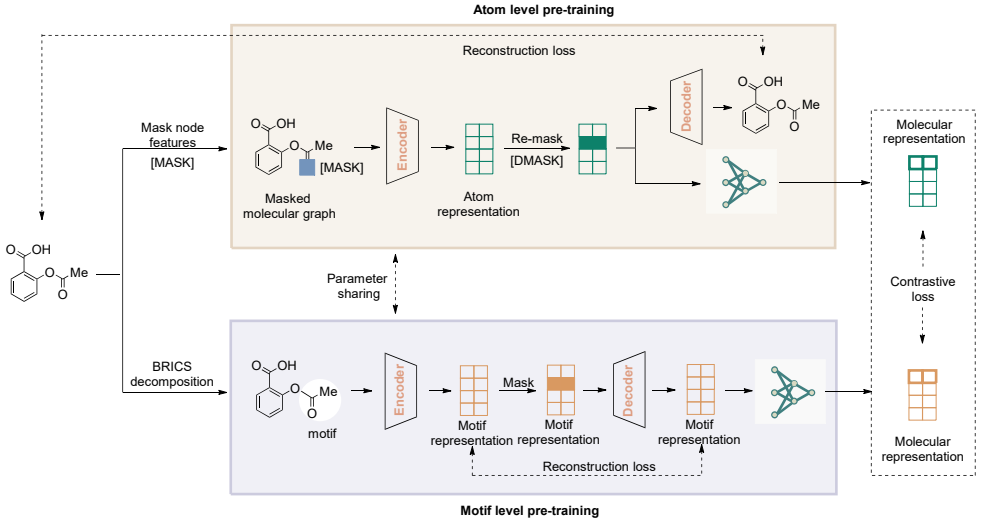

Fig. S1 Stage I: Pre-training on unlabeled molecular structures. The structure information on atom-level and motif-level is extracted through dual-level masking scheme. Additionally, we employ contrastive learning on dual-level features to further improve the diversity and robustness of molecular representations.

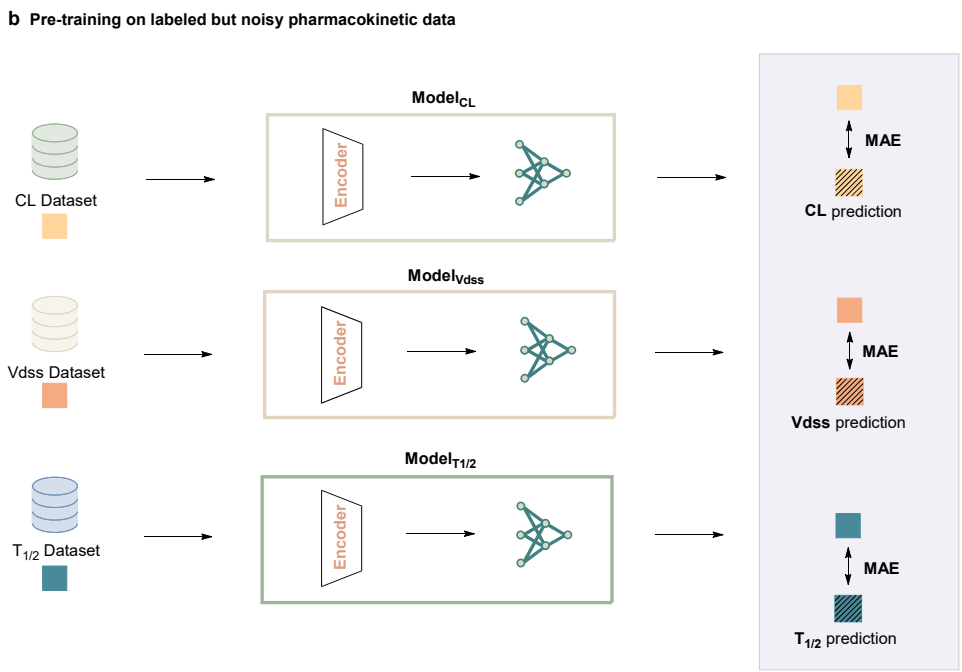

Fig. S2 Stage II: Pre-training on labeled but noisy pharmacokinetic data. We perform pre-training on three pharmacokinetic tasks, namely CL, Vdss, and $T_{1/2}$. We use the graph encoder from Stage I as the backbone for the three tasks.

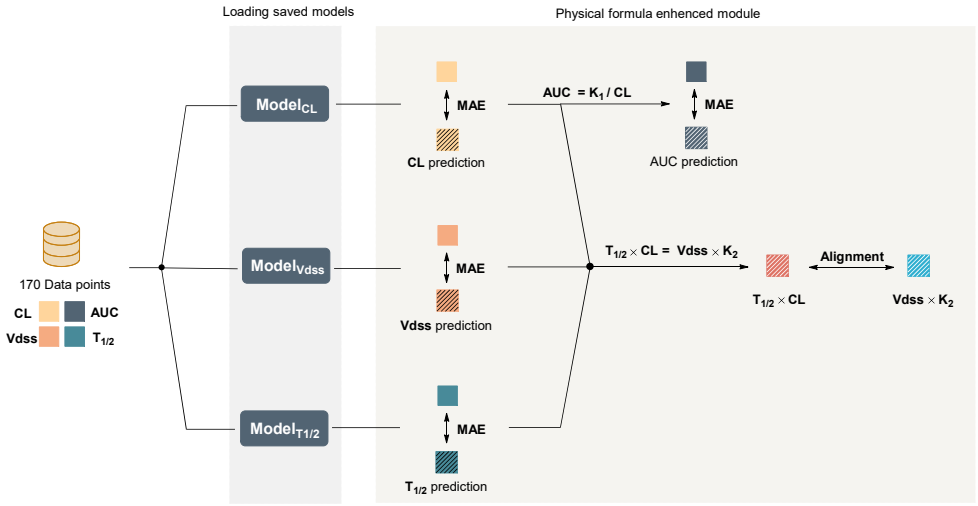

Fig. S3 Stage III: Physical formula enhanced multi-task learning. By incorporating physical formula constraints into the neural network, we enhance the knowledge transfer and objective alignment across different tasks.

2 Dataset Details

We select data of rats from the ChEMBL database [1] to build our dataset for the pharmacokinetic parameters. In stage III, since the AUC task is particularly susceptible to experimental conditions, we add two additional constraints besides animal type, that is, intravenous injection as the administration method and 1mg/kg as the dosage. Finally, we obtain 215 data points that have labels for all pharmacokinetic parameters. Data volume and screening criteria for each stage are summarized in Table 1.

Table 1 Dataset details for each stage[a].

Stage	Data volume	Label type
Stage I	904,816	-
Stage II	9,982	CL
	4,771	Vdss
	10,126	$T_{1/2}$
Stage III	215	AUC, CL, Vdss and $T_{1/2}$

[a] All dataset are collected from ChEMBL database. In the stage I, we filter data based on IC50; in the stage II, we filter data based on rat; and in the stage III, we filter data based on intravenous injection at a concentration of 1 mg/kg in rats."

3 Baseline methods

We compare PEMAL with four other methods: (1) Gaussian Process (GP) [2] (2) Random Forest (RF) [3], (3) Multilayer Perception (MLP), (4) Graph Isomorphism Network (GIN) [4]. GP is a commonly applied non-parametric Bayesian model for regression and classification tasks, which is capable of providing uncertainty estimates for predictions. RF is an ensemble learning method for regression by constructing a number of decision trees. We directly invoke the integrated scikit-learn [5] to implement GP and RF. MLP, as one of the simplest deep neural networks, possesses a relatively strong expressive capability. GIN excels in graph representation learning, particularly in tasks such as graph classification and isomorphism testing, where the complex relationships between graph topology and node features can be effectively learned.

4 The PMEAL: Details and hyperparameters

In Table 2, we show the hyperparameters for the pre-training on stage I. The configurations of CL, Vdss, and $T_{1/2}$ for stage II are shown in Table 3, Table 4, and Table 5, respectively. The hyperparameters of stage III are summarized in Table 6.

Table 2 Hyperparameter space considered for the PAMEL with the pre-training on stage I.

Hyperparameter	Explored values
Number of GNN layers	5
Feature dimension	300
Batch size	64
Learning rate	0.001
Optimizer	Adam
Weight decay	0
Dropout	0.0
Layer-normalization	False
Reconstruction loss function	Scaled cosine error
Contrastive loss function	InfoNCE loss
Atom mask ratio	0.75
Motif mask ratio	0.25
Mask SCE ratio	0.5
Contrastive ratio	0.5

Table 3 Hyperparameter space considered for the CL model with the pre-training on stage II.

Hyperparameter	Explored values
Number of GNN layers	5
Feature dimension	300
Batch size	64
Learning rate	0.0001
Optimizer	Adam
Weight decay	1e-10
Dropout	0.0
Layer-normalization	True

Table 4 Hyperparameter space considered for the Vdss model with the pre-training on stage II.

Hyperparameter	Explored values
Number of GNN layers	5
Feature dimension	300
Batch size	64
Learning rate	0.0005
Optimizer	Adam
Weight decay	1e-10
Dropout	0.0
Layer-normalization	True

Table 5 Hyperparameter space considered for the $T_{1/2}$ model with the pre-training on stage II.

Hyperparameter	Explored values
Number of GNN layers	5
Feature dimension	300
Batch size	64
Learning rate	0.0001
Optimizer	Adam
Weight decay	1e-10
Dropout	0.0
Layer-normalization	True

Table 6 Hyperparameter space with the stage III: physical formula enhanced multi-task learning.

Hyperparameter	Explored values
Number of GNN layers	5
Feature dimension	300
Batch size	32
Learning rate	0.0001
Optimizer	Adam
Weight decay	1e-9
Dropout	0.0
Layer-normalization	True
K_1	16,846
K_2	17.71
CL learning rate scale	1.0
Vdss learning rate scale	1.0
$T_{1/2}$ learning rate scale	$1.0 \sim 2.2$

References

[1] Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E., *et al.*: The chembl database in 2017. Nucleic acids research **45**(D1), 945–954 (2017)

[2] MacKay, D.J., *et al.*: Introduction to gaussian processes. NATO ASI series F computer and systems sciences **168**, 133–166 (1998)

[3] Biau, G., Scornet, E.: A random forest guided tour. Test **25**, 197–227 (2016)

[4] Xu, K., Hu, W., Leskovec, J., Jegelka, S.: How powerful are graph neural networks? In: 7th International Conference on Learning Representations, ICLR (2019)

[5] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.*: Scikit-learn: Machine learning in python. the Journal of machine Learning research **12**, 2825–2830 (2011)