nature portfolio

https://doi.org/10.1038/s42256-021-00301-6

Supplementary information

Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations

In the format provided by the authors and unedited

Out-of-the-Box Deep Learning Prediction of Pharmaceutical Properties by Broadly Learned Knowledge-Based Molecular Representations

Supplementary Information

Wan Xiang Shen[1,2], Xian Zeng[3], Feng Zhu[4], Ya li Wang[2], Chu Qin[2], Ying Tan[1,5], Yu Yang Jiang[1*] and Yu Zong Chen[2*]

[1] The State Key Laboratory of Chemical Oncogenomics, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, 518055, P. R. China

[2] Bioinformatics and Drug Design Group, Department of Pharmacy, and Center for Computational Science and Engineering, National University of Singapore, 117543, Singapore.

[3] Department of Biological Medicines & Shanghai Engineering Research Center of Immunotherapeutics, Fudan University School of Pharmacy, Shanghai 201203, P. R. China.

[4] College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, P. R. China.

[5] Shenzhen Kivita Innovative Drug Discovery Institute, Shenzhen, 518110, P. R. China

*Correspondence: Yu Zong Chen (phacyz@nus.edu.sg) and Yu Yang Jiang (jiangyy@sz.tsinghua.edu.cn).

CONTENTS

Part1: Supplementary Figures

Part2: Supplementary Tables

Table S7 | The top 10 important input-features of the training set of a MolMapNet-D solubility prediction model.

Table S8 | List of the fingerprint features that annotated in the Extended Data Fig. 9

Table-S9 | Detailed information about the datasets, splits, and code repositories

Part3: Supplementary Methods

1. Benchmark dataset splitting, modeling, and robustness test on 10 different random seeds
2. kNN and MolMapNet-F modelling on the fingerprints
3. BACE external test dataset collection and chemical space analysis
4. The feature importance calculation
5. The highlighting of atom and bond importance

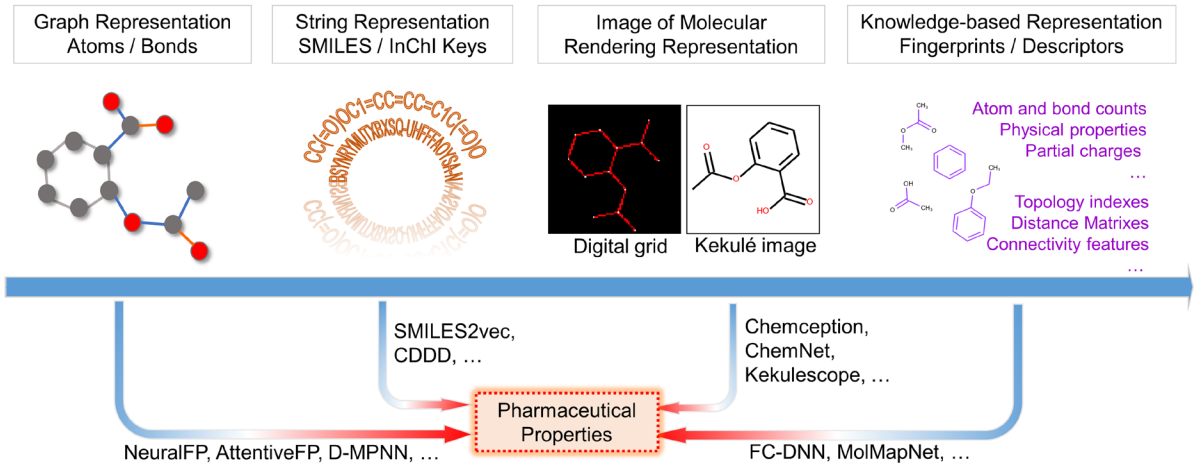

Fig. S1 | Example of the four types of molecular representations in molecule deep learning of pharmaceutical properties. In graph representations, the atoms and bonds are regarded as nodes and edges, and the aggregated node features are used by the graph convolutional network (GCN) models (e.g. NeuralFP[1], AttentiveFP[2] (attention mechanism), D-MPNN[3]) for pharmaceutical learning tasks. The string representations such as SMILES are typically learned by recurrent neural networks (RNNs) (e.g. SMILES2vec[4] and CDDD[5]) for generating vector-based molecular embeddings. The pixelized images of molecular rendering representations in digital grid or Kekulé image format are used as inputs of CNN models (e.g. Chemception[6], ChemNet[7] and Kekulescope[7]). The knowledge-based representations are expert domain knowledge-based molecule descriptors/fingerprints, which are inputs of the fully connected deep neural networks (FC-DNNs) or convolutional neural networks (CNNs) for pharmaceutical learning tasks.

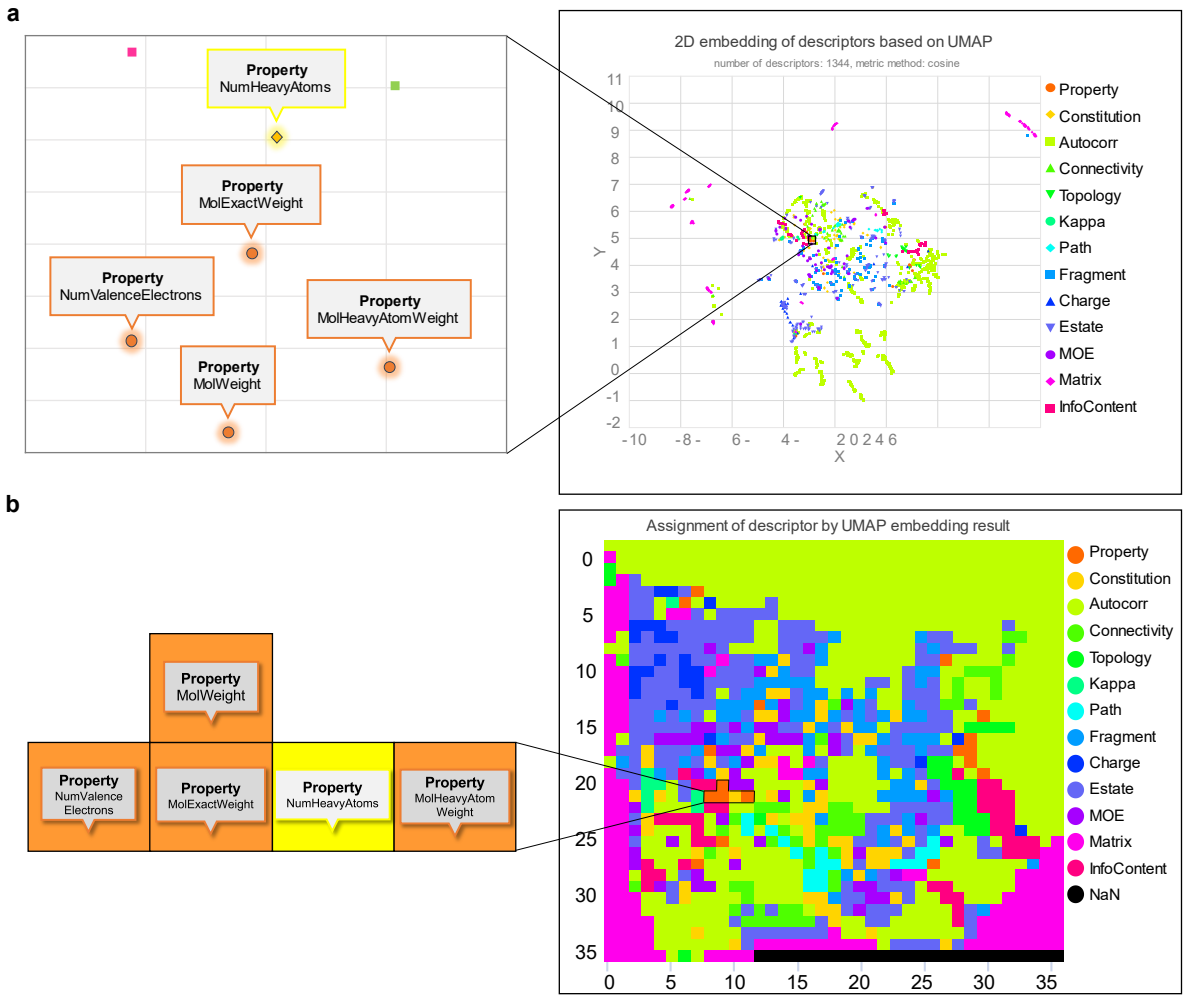

Fig. S2 | The scatter and grid distribution for molecular descriptors feature map. a, the embedding map generated by UMAP[8] of the 13 classes of molecular descriptors. **b,** the grid assignment of the UMAP embedding of the 13 classes of molecular descriptors.

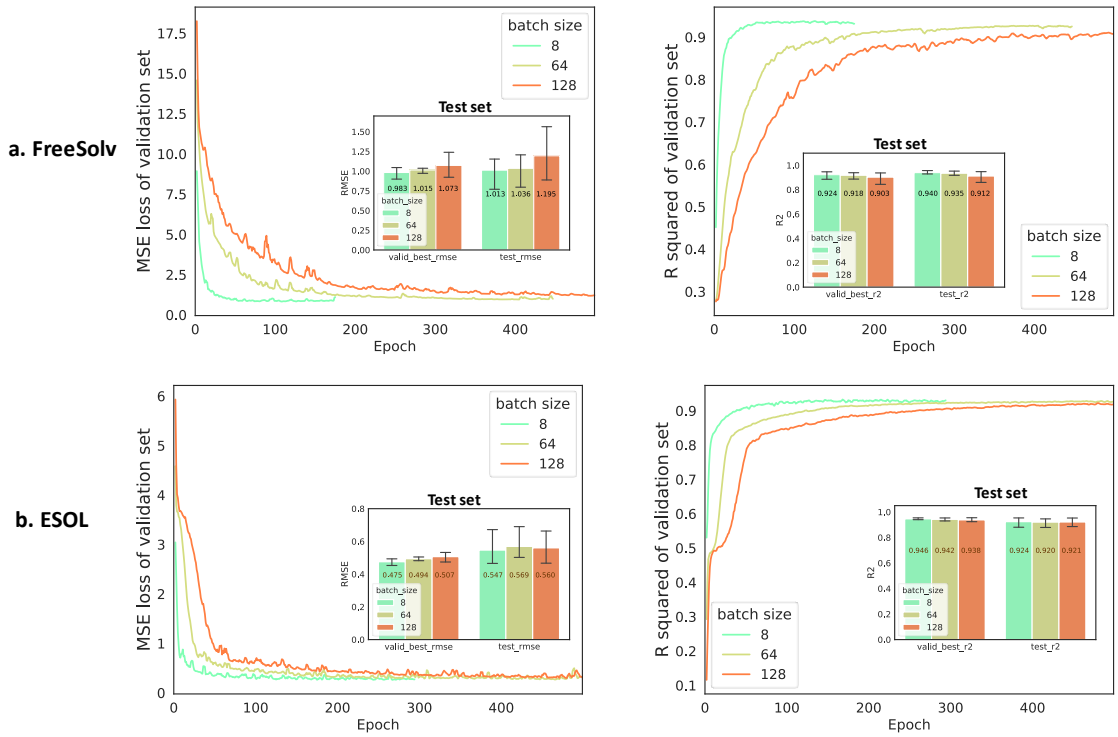

Fig. S3 | The batch size optimization of the training-control parameters of MolMapNet-B on the two regression data sets. a, the validation set performance during training and the final test set performance on FreeSolv (left: RMSE metric, right: R-squared matric). **b,** the validation set performance during training and the final test set performance ESOL. Smaller batch sizes showed a faster convergence than larger batch sizes for both data sets, leading to a better performance, the error bars represent standard error of the mean.

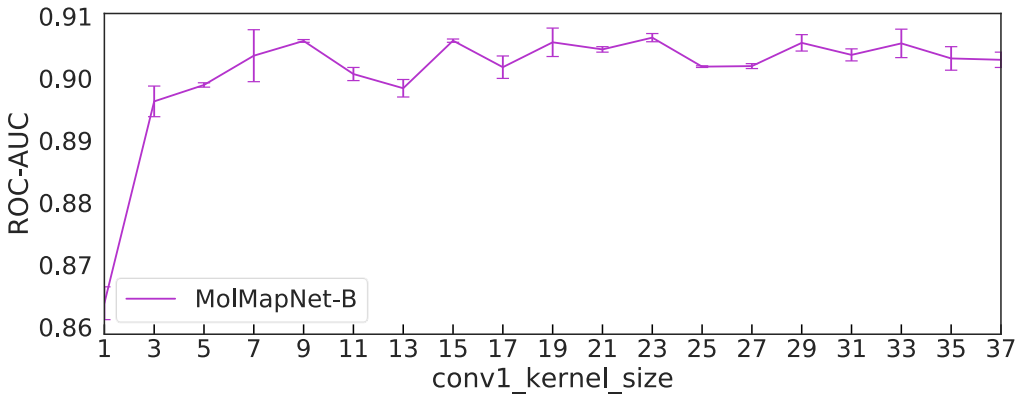

Fig. S4 | The impact of the kernel size of first convolution layer on the predictive performance of MolMapNet-B BACE classification model. The data set is under the AttentiveFP data-splits, the error bars represent standard error of the mean, the default kernel size in MolMapNet-B is 13

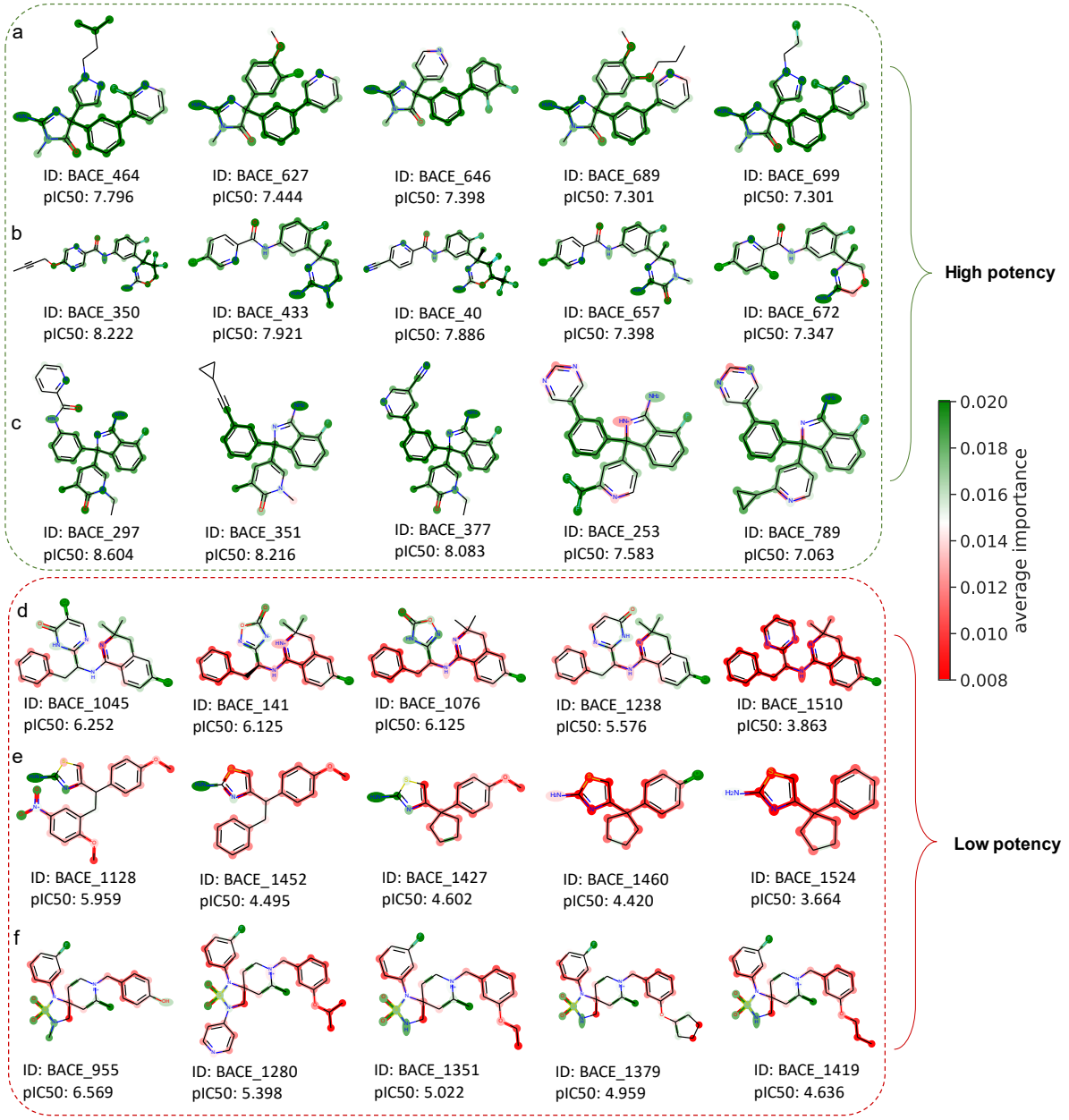

Fig. S5 | The average importance of the atoms and bonds of the BACE inhibitors of the other six molecular scaffolds in the BACE benchmark dataset. The compounds are color-highlighted based on the presence of the top50 important features (green color indicates higher average importance, red color lower importance), and their bioactivity in pIC50 values are provided. Compounds with higher portions of the important features (green) tend to have higher activity values (high potency).

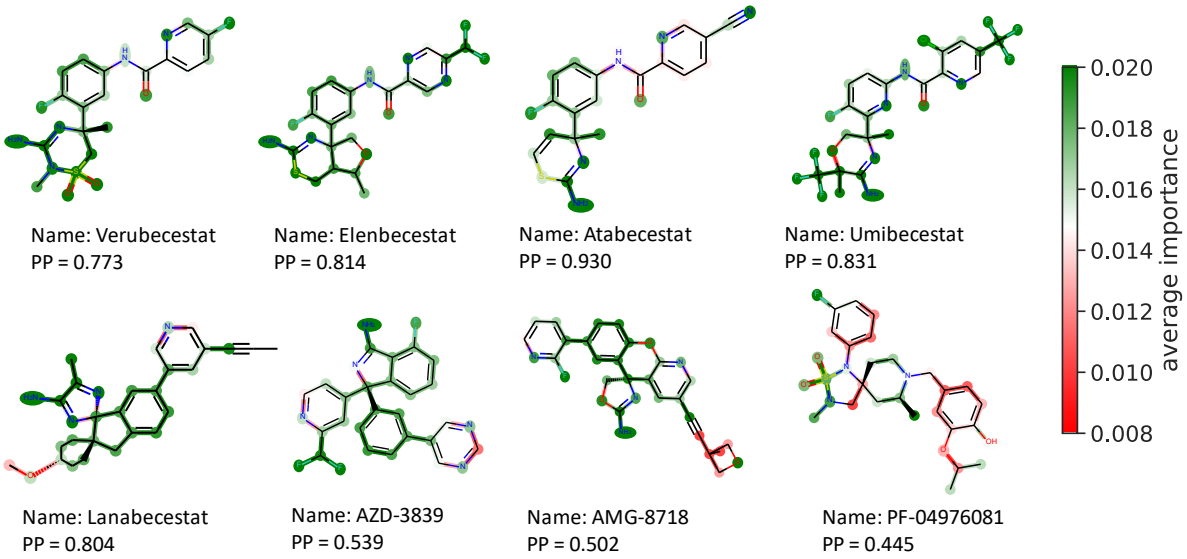

Fig. S6 | The average importance of the atoms and bonds of the 8 clinical trial drugs and the predicted probability (PP) as high potency inhibitors. All but the last drug are with overwhelmingly high portion of importance features (green) and PP > 0.5 values (indicative of potent inhibitor) by the BACE MolMapNet-F classification model.

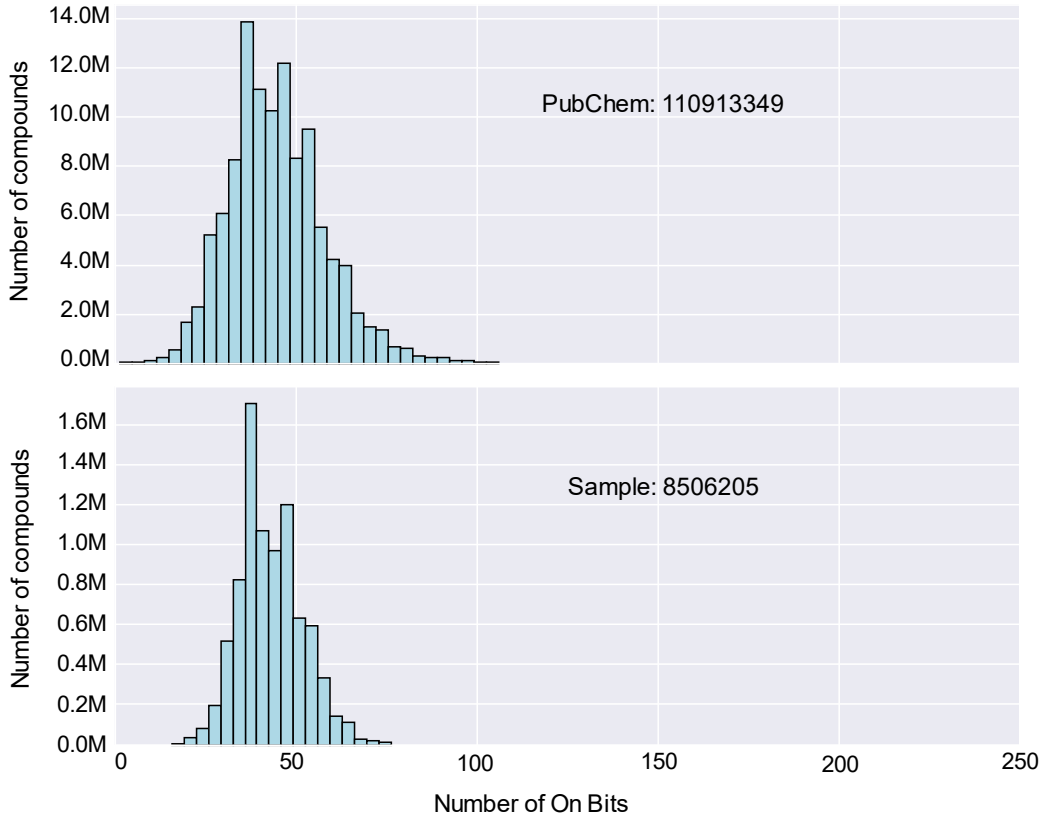

Fig. S7 | The distribution of NumOnBits for the original PubChem compounds and the sampled compounds using stratified sampling strategy. Their NumOnBits are calculated from the 2048-bit Morgan fingerprint (r=2, ECFP4-like).

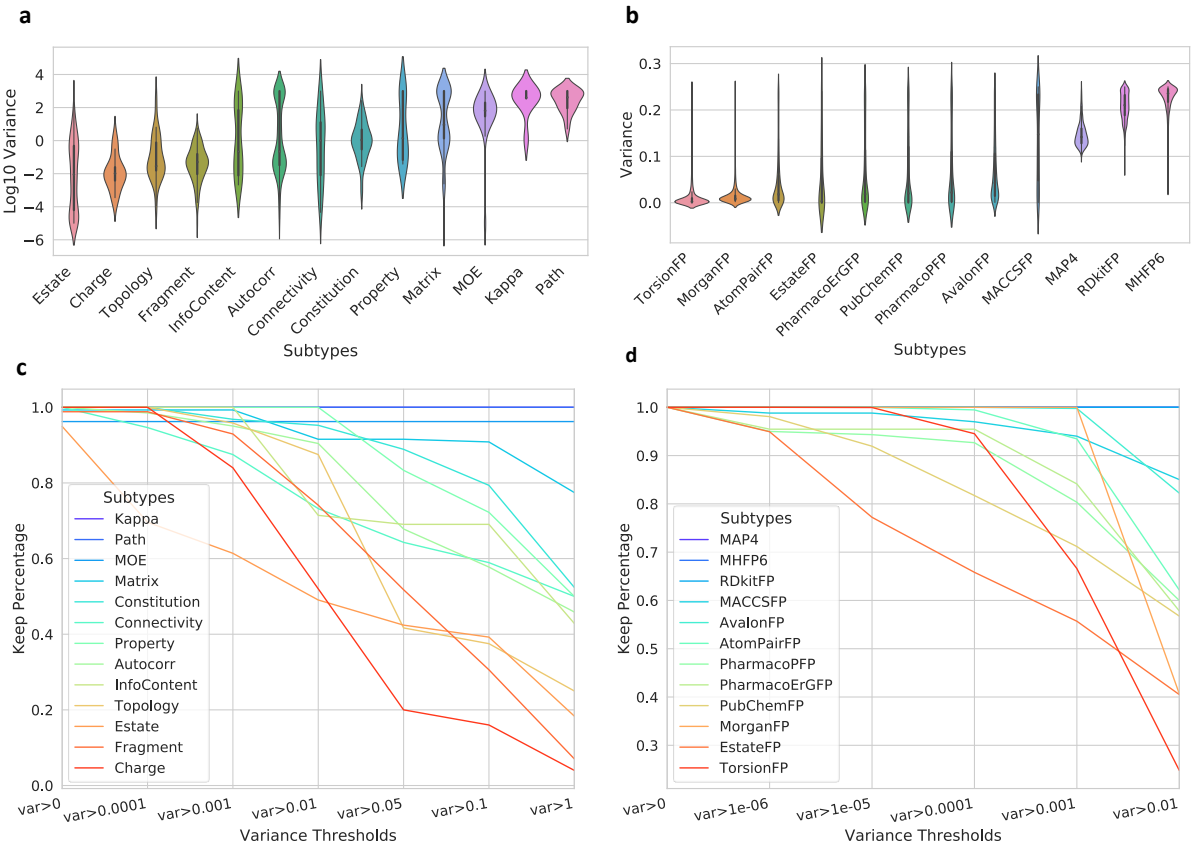

Fig. S8 | The variance distribution and the keep percentage of molecular features by using different variance thresholds. a, the variance (log10 value) distribution of the 13 classes of molecular descriptors. **b,** The variance distribution of the 9 sets of molecular fingerprints. **c** and **d,** the keep percentage of the descriptor sets and fingerprint sets respectively by using different variance thresholds. Some of these molecular features are of very low variance with near-zero values, which are less important for predicting pharmaceutical properties and more difficult for estimating the distances among them. Therefore, a variance threshold was applied to filter out these very low variance features. In MolMapNet, the default variance threshold is set to 1e-4. The higher the threshold, the more features are filtered out. Under this threshold, the average keep ratio is >90% for both molecular descriptors and fingerprint features.

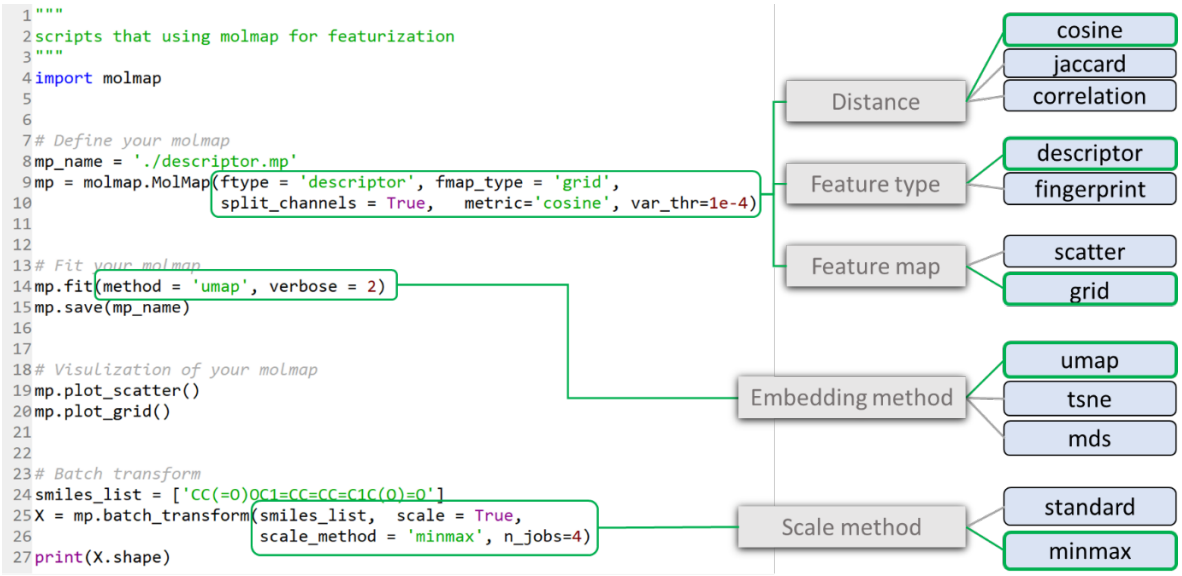

Fig. S9 | The example code for MolMap featurization, multiple methods are supported in some steps. The "split_channels" parameter is used for splitting the input-features into separate feature maps by the groups of the input-features (e.g. one or more classes of descriptors or sets of fingerprints), the "var_thr" parameter is used for filtering out the very low variance features, and the "n_jobs" parameter is used for parallel transformation. The feature maps can be normalized by a user-selected scaling method ("minmax" or "standard"). By providing an user-friendly MolMap package, users can easily generate their self-defined molecule feature maps for featurization and for deep learning of molecular and pharmaceutical properties by auto feature extraction, transforming and scaling.

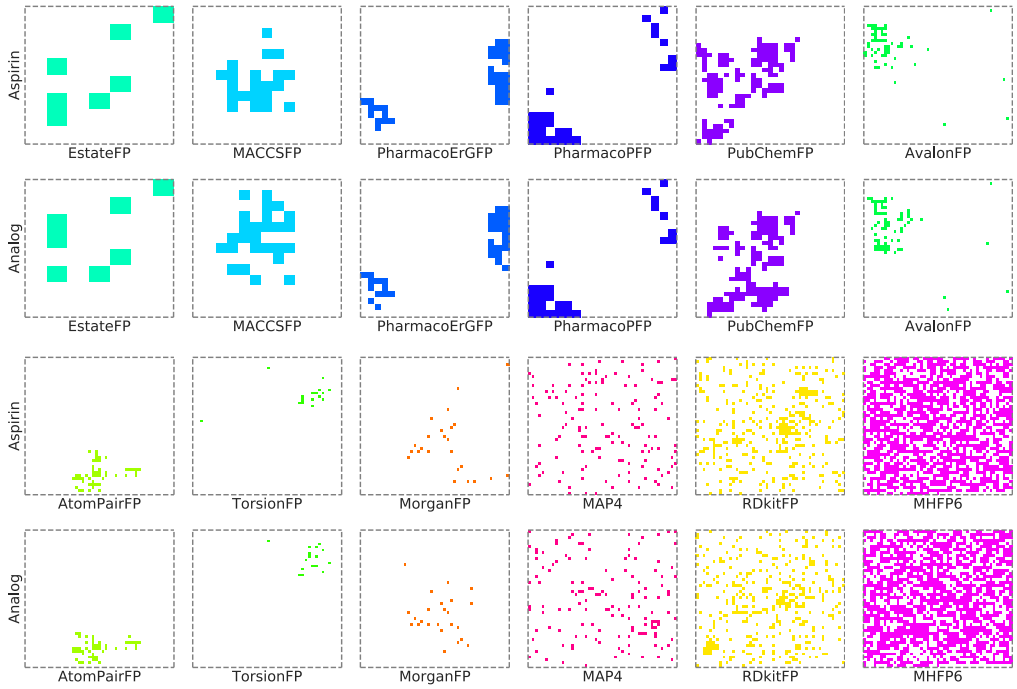

Fig. S10 | The 2D feature maps for aspirin and its analog N-acetylanthranilic acid that are generated by MolMap using 12 different sets of fingerprints. The MorganFP (ECFP4-liked) is calculated using radius=2, other settings on these fingerprints' calculation are provided in **Table-S3.**

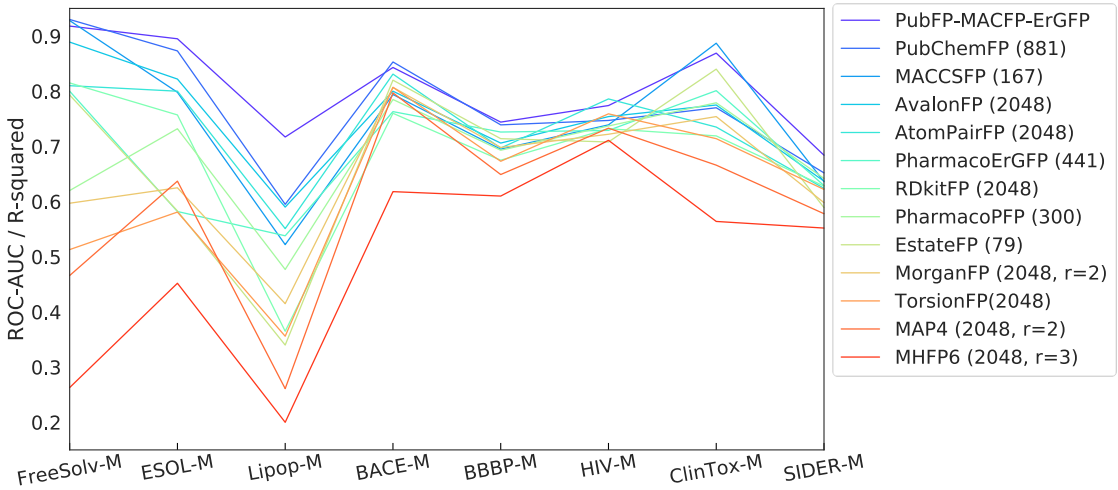

Fig. S11 | The average performance of the single-path MolMapNet-F using the different sets of fingerprints on the 5 benchmark datasets. The model is evaluated on total 12 different sets and a combination of three fingerprint sets PubFP-MACFP-ErGFP (PubChemFP, MACSSFP, and PharmacoErG)**.** The y-axis presents the R-square values for the 3 regression tasks (FreeSolv, ESOL, Lipop) or the ROC-AUC values for the 5 classification tasks, all datasets are split by MoleculeNet data-splits method.

Part 2: Supplementary Tables

Table-S1 | Comparison of the published works in the prediction of pharmaceutical properties. (**CDDD**: Continuous and Data-Driven Descriptors)

model	input feature type	representation	model architecture
NeuralFP[1], **AttentiveFP**[2] **D-MPNN**[3]	atom and bond features	graph	GCN or GAT
SMILES2vec[4], **CDDD**[5]	smiles	string	RNN/AENN
Chemception[6], ChemNet[7], **KekuleScope**[9]	grid or Kekulé image pixels	molecular structure graph	CNN
MolMapNet	descriptors/fingerprints	knowledge-based	CNN

Table-S2 | 13 classes of molecular descriptors covered by MolMap package

Class	# of features	Class	# of features
Autocorr	606	InfoContent	42
Estate	316	Charge	25
Matrix	142	Topology	24
Fragment	85	Property	18
Constitution	63	Path	18
Connectivity	56	Kappa	8
MOE	53	**Total**	**1456**

Table-S3 | 12 sets of molecular fingerprints covered by MolMap Package

Set	Number of Fingerprint Features in Default Settings
EstateFP[10]	79 bits
MACCSFP	167 bits, (1 + 166, Bit 0 is a placeholder)
PharmacoErGFP[11]	441 bits, minPath = 1, maxPath = 21
PharmacoPFP[12]	300 bits, minPointCount = 2, maxPointCount = 2
PubChemFP	881 bits (v1.3)
AvalonFP[13]	nBits = 2048
AtomPairFP	nBits = 2048, minLength=1, maxLength=30
TorsionFP	nBits = 2048, targetSize = 4
MorganFP (ECFP-like)	nBits = 2048, radius=2 (folded)
RDkitFP (DaylightFP-like)[14]	nBits = 2048, minPath=1, maxPath=7 (folded)
MHFP[15]	nBits = 2048, radius = 3 (folded)
MAP4[16]	nBits = 2048, radius = 2 (folded)
Total number of bits	**16204**

Table-S4 | Hyperparamters and the Out-of-the-Box Settings in MolMapNet

Hyperparameters	Suggest Options	Out-of-the-Box Setting
Featurization Parameters		
input feature maps	{'descriptor', 'fingerprint', 'both'}	'both'
metrics for feature point distance	{'cosine', 'correlation', 'jaccard'}	'cosine'
method for feature point embedding	{ 'ump', 'tsne', 'mds'}	'umap'
embedding parameters	hyperparameters in embedding method such as: n_neighbors, min_dist, max_iter, perplexity, etc.	n_neighbors = 30; min_dist = 0.1
split channels	{True, False}	True
scale method	{'minmax', 'standard'}	'minmax'
Network Architecture Parameters		
model path	single or double path, depends on the input feature maps	double path
conv1_kernel_size	Odd number, 1~37	13
# of dense layers	1-3 layers; depends on double path or single path	3
# of units per dense layer	pyramidal, depends on the number of the outputs	[256, 128, 32] *
activation function in dense layers	{'relu', 'tanh'}	'relu'
activation function in last layer	regression: linear; classification: sigmoid	regression: linear; classification: sigmoid
Training-Control Parameters		
optimizer	{Adam, SGD, etc.}	Adam
learning rate	1e-2, 1e-3, 1e-4, 1e-5	1e-4
learning rate decay	0.0~0.1	0.0
dropout rate	0.0~0.5	0.0
weight decay	0.0~0.5	0.0
batch size	1~1024	128 #
loss	regression: MSE/MAE; classification: (weighted) cross entropy	regression: MSE; classification: (weighted) cross entropy
monitor for early stopping[↓]	performance of the validation set	loss/metrics of the validation set

***** For some of the multi-tasks such as MUV, PCBA, ChEMBL, Tox21, SIDER, ToxCast, the outputs are more than one unit, so the dense layers and units are set differently, details are provided in the **Source_data_to_Extended_Data_Fig.3.xlsx** file. # For the regression tasks of low-data cases, a smaller batch size is recommended for better convergence. A "patient" parameter is used for early stopping, i.e., if the performance of validation set (the performance of validation set comes from the callbacks of each epoch) shows no improvement in the next patience (50 epochs), the training process will be terminated. For each epoch, the current model is compared with the previously saved model, and the best model is saved automatically. In the end, the best model is saved as the optimized model.

Table-S5 | Performance of the optimized MolMapNet-B models in comparison with the state-of-the-art graph-based models and MolMapNet-B out-of-the-box models. The models are tested on the 2 physicochemical and 6 bioactivity properties prediction tasks, **t**he bold indicates the best performing model without the optimized MolMapNet-B models. The underlined bold indicates the cases of the optimized MolMapNet-B models outperforming all other models. The results marked by the red * label indicates significantly improved performance of the optimized MolMapNet-B models over all other models

Data Class	**Dataset**	**Task Metric**	**Performance**				
			MoleculeNet[17]	**Chemprop[3]**	**AttentiveFP[2]**	**MolMapNet-B**	
			(GCN best)	**(DMPNN)**		**OOTB**	**Optimized**
Physico-chemical	ESOL	RMSE	0.580 (MPNN)	**0.555**		0.575	**<u>0.544</u>**
					0.486	0.543	**<u>0.512</u>**
	FreeSolv	RMSE	1.150 (MPNN)	**1.075**		1.155	**<u>0.916</u>***
					0.773	0.994	**<u>0.812</u>***
Bio-activity	Malaria	RMSE			1.077	**1.011**	**<u>1.008</u>**
	BACE	ROC_AUC	0.806 (Weave)	N.A.		**0.849**	**<u>0.854</u>**
					0.856	**0.881**	**<u>0.891</u>**
	HIV	ROC_AUC	0.763 (GC)	0.776		**0.777**	**<u>0.788</u>**
					0.848	**0.865**	**<u>0.87</u>**
	MUV	PRC_AUC	**0.109 (**Weave)	0.041		0.096	**<u>0.158</u>***
	PCBA	PRC_AUC	0.136 (GC)	**0.335**		0.276	0.276
	ChEMBL	ROC_AUC		0.739		**0.75**	**<u>0.766</u>**

Table S6 | The predictive performance of the 4 BACE classification models on the external ChEMBL dataset. This dataset contains BACE inhibitors of novel molecular scaffolds extracted from ChEMBL, the extraction method is described in the **Supplementary method 3,** the kNN model was built by the same fingerprints as MolMapNet-F.

Evaluation Set	Model	Sensitivity	Specificity	ROC-AUC	TP	TN	FN	FP
Validation set (N=151: 77, 74)	kNN	0.73	0.86	0.84	56	64	21	10
	AttentiveFP	0.77	0.73	0.82	59	54	18	20
	DMPNN	0.57	0.91	0.86	44	67	33	7
	MolMapNet-F	**0.78**	**0.84**	**0.88**	60	62	17	12
Test set, (N=152: 50, 102)	kNN	0.56	0.92	0.87	28	94	22	8
	AttentiveFP	0.74	0.84	0.84	37	86	13	16
	DMPNN	0.66	0.88	0.86	33	90	17	12
	MolMapNet-F	**0.84**	**0.87**	**0.89**	42	89	8	13
novel ChEMBL set (N=395: 216, 179)	kNN	0.24	0.90	0.63	52	161	164	18
	AttentiveFP	0.63	0.63	0.70	137	113	79	66
	DMPNN	0.48	0.81	0.72	103	145	113	34
	MolMapNet-F	**0.70**	0.84	**0.79**	151	150	65	29
common ChEMBL set (N=5325: 2077, 3248)	kNN	0.75	0.65	0.76	1551	2126	526	1122
	AttentiveFP	0.67	0.78	0.81	1385	2543	692	705
	DMPNN	0.52	0.84	0.79	1084	2738	993	510
	MolMapNet-F	**0.85**	**0.81**	**0.87**	1771	2629	306	619

(N = total number of samples: # of inhibitors, # of non-inhibitors), **TP**: True Positive, **TN**: True Negative, **FP**: False Positive, **FN**: False Negative

Table S7 | The top 10 important input-features of the training set of a MolMapNet-D solubility prediction model trained on the ESOL dataset using the AttentiveFP data-split.

rank	feature point	subtype	description	Score
1	MAXdsssP	Estate	maximum atom-type E-State: ->P=	0.313
2	MolQedWeightsNone	Drug-likeness	QED descriptor using unit weights	0.259
3	MAXsssCH	Estate	maximum atom-type E-State: >CH-	0.202
4	NChargeMean	Charge	average negative charge	0.195
5	AXp-0d	Connectivity	averaged Chi path weighted by sigma electrons	0.163
6	MINdsssP	Estate	minimum atom-type E-State: ->P=	0.158
7	VE2_A	Matrix	average eigenvector coefficient sum from adjacency matrix	0.144
8	MINsCH3	Estate	minimum atom-type E-State: -CH3	0.141
9	VE2_Dzare	Topological index	Barysz matrix weighted by electronegativity	0.134
10	VE2_DzZ	Topological index	Barysz matrix weighted by atomic number	0.120

Table S8 | List of the fingerprint features that annotated in the Extended Data Fig. 9

Group	FP id	FP smarts	FP loc (x, y)	FP imp	FP rank
Group1	PharmacoErGFP286	('Positive', 'Hydrophobic', 14)	(03, 32)	0.018202	29
Group1	PharmacoErGFP144	('Acceptor', 'Acceptor', 19)	(03, 33)	0.010709	76
Group1	PharmacoErGFP287	('Positive', 'Hydrophobic', 15)	(04, 30)	0.027124	10
Group1	PharmacoErGFP283	('Positive', 'Hydrophobic', 11)	(04, 34)	0.028608	9
Group1	PharmacoErGFP149	('Positive', 'Acceptor', 3)	(05, 26)	0.036896	6
Group1	PharmacoErGFP289	('Positive', 'Hydrophobic', 17)	(05, 27)	0.017354	35
Group1	PharmacoErGFP148	('Positive', 'Acceptor', 2)	(05, 29)	0.046328	3
Group1	PharmacoErGFP146	('Acceptor', 'Acceptor', 21)	(05, 31)	0.020916	22
Group1	PharmacoErGFP44	('Positive', 'Donor', 3)	(06, 32)	0.017269	36
Group1	PharmacoErGFP42	('Positive', 'Donor', 1)	(07, 33)	0.010146	80
Group1	PharmacoErGFP41	('Donor', 'Acceptor', 21)	(07, 34)	0.013898	54
Group2	PharmacoErGFP26	('Donor', 'Acceptor', 6)	(29, 09)	0.023202	16
Group2	PubChemFP16	('[N]', 3)	(30, 08)	0.013381	58
Group2	PharmacoErGFP5	('Donor', 'Donor', 6)	(30, 09)	0.017973	31
Group2	PharmacoErGFP24	('Donor', 'Acceptor', 4)	(30, 10)	0.039897	5
Group2	PharmacoErGFP4	('Donor', 'Donor', 5)	(31, 08)	0.012499	63
Group2	MACCSFP84	('[NH2]', 0)	(31, 09)	0.044342	4
Group2	PharmacoErGFP104	('Donor', 'Hydrophobic', 21)	(31, 10)	0.026633	12
Group2	MACCSFP53	('[!#6;!#1;!H0]~*~*~*~[!#6;!#1;!H0]', 0)	(32, 08)	0.013425	56
Group2	MACCSFP131	('[!#6;!#1;!H0]', 1)	(32, 09)	0.025407	15
Group2	PharmacoErGFP103	('Donor', 'Hydrophobic', 20)	(32, 10)	0.025697	14
Group3	PubChemFP797	('[#6][#6]1[#6][#6]([#6])[#6][#6][#6]1', 0)	(13, 29)	0.051609	1
Group3	PubChemFP696	('[#6]-,:[#6]-,:[#6]-,:[#6]-,:[#6]-,:[#6]-,:[#6]-,:[#6]', 0)	(14, 29)	0.017727	33
Group3	PubChemFP697	('[#6]-,:[#6]-,:[#6]-,:[#6]-,:[#6]-,:[#6](-,:[#6])-,:[#6]', 0)	(15, 29)	0.022481	17
Group3	PubChemFP712	('[#6]-,:[#6](-,:[#6])-,:[#6](-,:[#6])-,:[#6]', 0)	(16, 29)	0.014196	53
Group3	PubChemFP734	('[#6]c1cc([#6])ccc1', 0)	(16, 31)	0.047897	2
Group4	PubChemFP364	(‘[#6](~[F])(:c)’, 0)	(26, 12)	0.012287	64

Group4	PubChemFP287	(‘[#6]~[F]’, 0)	(26, 13)	0.016003	42
Group4	MACCSFP42	(‘F’, 0)	(26, 14)	0.022456	18
Group4	PubChemFP363	(‘[#6](~[F])(~[F])’, 0)	(26, 15)	0.007065	112
Group4	MACCSFP87	(‘[F,Cl,Br,I]!@*@*’, 0)	(27, 12)	0.017708	34
Group4	MACCSFP107	('[F,Cl,Br,I]~*(~*)~*’, 0)	(27, 13)	0.030806	7
Group4	MACCSFP134	(‘[F,Cl,Br,I]’, 0)	(27, 14)	0.029291	8
Group5	MACCSFP110	('[#7]~[#6]~[#8]', 0)	(29, 26)	0.016884	37
Group5	MACCSFP92	('[#8]~[#6](~[#7])~[#6]', 0)	(29, 27)	0.014277	51
Group5	PubChemFP536	('[#8]=,:[#6]-,:[#6]-,:[#7]', 0)	(30, 26)	0.017935	32
Group5	PubChemFP451	('[#6](-,:[#7])(=,:[#8])', 0)	(30, 27)	0.011341	69
Group5	MACCSFP154	('[#6]=[#8]', 0)	(31, 27)	0.020061	23
Group5	PubChemFP439	('[#6](-,:[#6])(-,:[#7])(=,:[#8])', 0)	(31, 29)	0.013384	57
Group5	PubChemFP420	('[#6]=,:[#8]', 0)	(32, 27)	0.011374	68
Group5	PubChemFP443	('[#6](-,:[#6])(=,:[#8])', 0)	(32, 29)	0.013113	59
Group5	PubChemFP579	('[#8]=,:[#6]-,:[#6]-,:[#6]-,:[#6]', 0)	(33, 28)	0.013083	61
Group5	PubChemFP685	('[#8]=,:[#6]-,:[#6]-,:[#6]-,:[#6]-,:[#7]', 0)	(34, 28)	0.015718	44
Group5	PubChemFP684	('[#8]=,:[#6]-,:[#6]-,:[#6]-,:[#6]-,:[#6]', 0)	(34, 29)	0.016723	40
Group5	PubChemFP692	('[#8]=,:[#6]-,:[#6]-,:[#6]-,:[#6]-,:[#6]-,:[#6]', 0)	(35, 29)	0.016716	41
Group5	PubChemFP704	('[#8]=,:[#6]-,:[#6]-,:[#6]-,:[#6]-,:[#6]-,:[#6]-,:[#6]', 0)	(35, 30)	0.008745	88
Group6	MACCSFP127	('*@*!@[#8]’, 1)	(16, 30)	0.018610	27
Group6	MACCSFP143	('*@*!@[#8]’, 0)	(17, 30)	0.026163	13

Table-S9 | Detailed information about the datasets, splits, and code repositories

Data Sets and Splits
CYP450: this dataset contains 16896 compounds against five main CYP450 isozymes: 1A2, 2C9, 2C19, 2D6, and 3A4. This data set is split **by assay ids** (AIDs) in previous paper[18]
LMC: this dataset contains 8755 compounds and their liver microsomal clearance in three species, human, rat and mouse. This data set is split by **random-split** method in previous paper[19]
A2780: this dataset contains 2255 compounds with pIC50 values for the ovarian carcinoma cell line, it is split by **random-split** method using a proportion of 0.7, 0.15, 0.15 for training, valid and test set[9].
CCRF-CEM: this dataset contains 3047 compounds with pIC50 values for T-cell leukemia, it is split by **random-split** method using a proportion of 0.7, 0.15, 0.15 for training, valid and test set[9].
DU-145: this dataset contains 2512 compounds with pIC50 values for the prostate carcinoma cells, it is split by **random-split** method using a proportion of 0.7, 0.15, 0.15 for training, valid and test set[9].
HCT-15 994: this dataset contains 994 compounds with pIC50 values for the colon adenocarcinoma cells, it is split by **random-split** method using a proportion of 0.7, 0.15, 0.15 for training, valid and test set[9].
KB 2731: this dataset contains 2731 compounds with pIC50 values for the squamous cell carcinoma, it is split by **random-split** method using a proportion of 0.7, 0.15, 0.15 for training, valid and test set[9].
LoVo 1120: this dataset contains 1120 compounds with pIC50 values for the colon adenocarcinoma cells, it is split by **random-split** method using a proportion of 0.7, 0.15, 0.15 for training, valid and test set[9].
PC-3: this dataset contains 4294 compounds with pIC50 values for the prostate carcinoma cells, it is split by **random-split** method using a proportion of 0.7, 0.15, 0.15 for training, valid and test set[9].
SK-OV-3: this dataset contains 1589 compounds with pIC50 values for the ovarian carcinoma cells, it is split by **random-split** method using a proportion of 0.7, 0.15, 0.15 for training, valid and test set[9].

ESOL: this dataset includes 1128 compounds and their experimental water solubility. This data set is split by **random-split** method in previous papers[17]
FreeSolv: this dataset contains 642 small molecules' experimental hydration free energy in water. This data set is split by **random-split** method in previous papers[17]
Lipop: this dataset has 4200 compounds and their corresponding experimental lipophilicity values. This data set is split by **random split** method. This data set is split by random-split method in previous papers[17]
BBBP: this dataset contains 2039 compounds with their binary permeability properties of Blood-brain barrier. This data set is split by **scaffold-split** method in previous papers[17]
PDBbind-F (full): this dataset contains 9880 compounds with their logKd/Ki binding affinity, this data set is split by **time-split** method in previous paper[17]
PDBbind-C (core): this subset of PDB-binding database contains 168 compounds with their logKd/Ki binding affinity, it is compiled as high-quality data sets of protein-ligand complexes for docking/scoring studies. This data set is split by **time-split** method in previous paper[17]
PDBbind-R (refined): this dataset contains 3040 compounds with their logKd/Ki binding affinity, it contains protein-ligand structures at higher resolution and excludes any complexes with IC50 affinity data only. This data set is split by **time-split** method in previous paper[17]
Malaria: this dataset includes 9998 compounds that experimentally measured EC50 values of a sulfide-resistant strain of Plasmodium falciparum, which is the source of malaria. This data set is split by **random-split** method (only split to train and validation set by a fraction of 0.2) in previous paper[2]
BACE: this dataset contains 1513 inhibitors with their binary inhibition labels for the target of BACE-1. This data set is split by **scaffold-split** method in previous papers[17]
HIV: this dataset contains 41127 compounds and their binary ability to inhibit HIV replication. This data set is split by **scaffold-split** method in previous papers[17]
Tox21: this dataset contains 7831 compounds and corresponding toxicity data against 12 targets. This data set is split by **random-split** method in previous papers[17]
SIDER: this dataset contains 1427 marketed drugs and their adverse drug reactions (ADR) against 27 System-Organs Class. This data set is split by **random-split** method in previous papers[17]

ClinTox: this dataset contains 1478 drugs or compounds; the labels are FDA approval status and clinical trial toxicity results. This data set is split by **random-split** method in previous papers[17]
PCBA: this dataset comes from PubChem Bioassay and contains 437929 compounds on 128 tasks that are related to biological activities, this data set is split by **random-split** method in previous papers[17] **MUV:** this dataset from PubChem Bioassay by applying a refined nearest neighbor analysis, it contains 17 challenging tasks for 93087 compounds and is specially designed for validation of virtual screening techniques. This data set is split by **random-split** method in previous papers[17]
ChEMBL: this dataset contains about 456331 compounds and more than 1310 assays. These assays correspond to a variety of target classes (e.g. enzymes, ion channels and receptors) and differ in size[20]. This dataset is split by scaffold-split method.[3]
Code Repos.
DeepChem(MoleculeNet)[21] : https://github.com/deepchem/deepchem/tree/2.2.0
AttentiveFP[2] repo: https://github.com/OpenDrugAI/AttentiveFP
Chemprop (D-MPNN)[3] repo: https://github.com/chemprop/chemprop
Kekulescope[9] repo: https://github.com/isidroc/kekulescope
ChemBench repo: https://github.com/shenwanxiang/ChemBench
MolMapNet GitHub repo: https://github.com/shenwanxiang/bidd-molmap CodeOcean repo: https://codeocean.com/capsule/2307823/tree

Part 3: Supplementary Methods

1. Benchmark dataset splitting, modeling, and robustness test on 10 different random seeds

To test the model robustness, we split the 12-benchmark datasets in 4 groups: **group1** contains 3 datasets (FreeSolv, ESOL, Malaria) of regression tasks using random split; **group2** includes 3 datasets (BACE, BBBP, HIV) of classification tasks using scaffold split; **group3** consists of 3 datasets (Tox21, ToxCast, SIDER) of multi-task classification tasks using random split; **group4** covers 3 datasets (MUV, PCBA, ChEMBL) of high-data and multi-task classification tasks using 10 different random seed: 2, 16, 32, 64, 128, 256, 512, 1024, 2048, 4096. The details about data splitting, modeling, and testing are as follows:

a) To split the dataset using random or scaffold method, we used the split tool from chemprop(https://github.com/chemprop/chemprop/blob/master/scripts/split_data.py) to split the benchmark data into train, validation, and test set by a proportion of 0.8, 0.1, 0.1.

b) To train the AttentiveFP model on different random seeds, we directly run their original code under the random seed generated training, validation and test set using their optimized parameters from: https://github.com/OpenDrugAI/AttentiveFP/tree/master/code.

c) To train the DMPNN model, we run the chemprop package from https://github.com/chemprop/chemprop/. We optimized the hyperparameters using the training set of each dataset and each split by the "chemprop_hyperopt" tool in their package, we subsequently built the model using training set by the optimized parameters and then used the validation set for the early stopping and best model selection among the epochs, and lastly, the best model was evaluated by the test set.

d) To train the MolMapNet Out-Of-The-Box (OOTB) model, we used the molmap package and the default parameters. Note that for the very large dataset PCBA and ChEMBL, we trained the model on the MolMapNet-Fingerprint only to save the computational costs.

The performance of the three models, AttentiveFP, DMPNN, and MMN-OOTB, under different random seed splits are shown in the line-plots of **Source Data to Extended Data Fig.1 and 2**.

2. kNN and MolMapNet-F modelling on the fingerprints

To test the performance of the kNN on 5 datasets (BACE, BBBP, HIV, ClinTox and SIDER) split by the MoleculeNet data-split method. The KNN regression and classification modelling

were conducted by sklearn (https://scikit-learn.org/stable/modules/classes.html#module-sklearn.neighbors) KNeighborsClassifier modules, the model was trained using the training set, the parameters n_neighbors and weight function was optimized using the grid-search method based on the performance of validation set, and the final model was evaluated on the test set. The input features for kNN models are the same as the MolMapNet-F, namely the three assemble fingerprints (PubChemFP, MACCSFP and PharmacoErGFP) PubFP-MACFP-ErGFP.

3. BACE external test dataset collection and chemical space analysis

a. **Data collection from ChEMBL.** The External BACE1 inhibitor data were extracted from ChEMBL. We downloaded the bio-assays data for the BACE1 target (ChEMBL4822), and kept the assay types such as IC50, KI50, Kd, etc. The pChEMBL value represents any one of the -Log (molar IC50, XC50, EC50, AC50, Ki, Kd or Potency).) [22], which was used as the prediction values or labels. Although there are more rigorous definitions of inhibitors and non-inhibitors, in this work we tentatively follow the original benchmark BACE dataset (according to their classification labels and pIC50 values, a pIC50 cutoff of 7 is used)[17] to divide the high potency inhibitors and low potency inhibitors based on their activity cut-off (pChEMBL value:7). The duplicates within the collected data and with respect to the BACE benchmark set were removed, leading to 5720 ChEMBL BACE compounds (3427 low potency inhibitors and 2293 high potency inhibitors). The structure and the pChEMBL values of these compounds are available in **Source_data_to_Extended_Data_Fig.6.xlsx**

b. Novel BACE external test dataset. We extracted 395 novel compounds from the 5720 ChEMBL BACE compounds with respect to the BACE benchmark set. To extract the novel compounds, we first mixed 5720 ChEMBL BACE compounds and the BACE benchmark compounds together, which were clustered by means of hierarchical clustering into 30 clusters using 2048-bit Morgan fingerprints (r = 2). We then selected the 22nd cluster that contains only ChEMBL BACE compounds. This cluster contains 395 compounds (216 high potency inhibitors and 179 low potency inhibitors), their average maximal pairwise Tanimoto similarity with respect to the BACE benchmark set is 0.372, suggesting that these 395 compounds are substantially novel in structures with respect to the BACE benchmark compounds. Therefore, these 395 compounds were regarded as the novel BACE external test dataset, and the rest of the 5325 ChEMBL BACE compounds were regarded as the common BACE external test dataset, the structures and activities of these compounds are available in **Source_data_to_Extended_Data_Fig.6.xlsx**

c. **BACE-1 clinical trial drugs.** We also collected 26 clinical trial drugs that are BACE-1 inhibitors from literatures[23,24] and commercial Cortellis Drug Discovery Intelligence (CDDI) database, the structure and activity of these drugs are available in **Source_data_to_Extended_Data_Fig.6.xlsx**

d. To explore the chemical space of the BACE data and the novel BACE external data, we applied the Tree-MAP(TMAP)[25] to visualize the compound distribution in the chemical space. Specifically, to generate the TMAP 2D embedding, the 1024 bits Morgan fingerprint (r=2, ECFP4-like) was used for the similarity calculation.

4. The feature importance calculation

The feature importance score $\boldsymbol{S}$ was calculated by the permutation algorithm[26,27] as follows:

Input: Trained model $\boldsymbol{f}$, feature matrix $\boldsymbol{X}$, target vector $\boldsymbol{y}$, error measure $\boldsymbol{L(y,f)}$. To estimate this error $\boldsymbol{L}$, the mean squared error is used for ESOL regression model and the log loss(cross-entropy) is used for the BACE classification model.

a) Estimate the original model error $\boldsymbol{e_{orig} = L(y, f(X))}$

b) For each feature $\boldsymbol{i = 1, \ldots, k}$ do:

Generate feature matrix $\boldsymbol{X_{perm}}$ by permuting feature $\boldsymbol{i}$ in the data $\boldsymbol{X}$. This breaks the association between feature $\boldsymbol{i}$ and true outcome $\boldsymbol{y}$.

Estimate error $\boldsymbol{e_{perm} = L(y, f(X_{perm}))}$ based on the predictions of the permuted data.

Calculate permutation feature importance score: $\boldsymbol{S_i = e_{perm} - e_{orig}}$

c) Sort features by descending feature importance score $\boldsymbol{S}$.

5. The highlighting of atom and bond importance

The highlighting the atom and bond importance of a given molecule is illustrated as follows: First, we selected the top50 important fingerprints in the BACE MolMapNet-F classification model, and then the importance on the atoms and bonds of this molecule was averaged by the following procedure, finally the green and red color indicate highly and moderately important substructure (i.e. higher and lower ranked in the top-50 fingerprints) respectively.

Input: A molecule $\boldsymbol{m}$ to highlight, the important fingerprints $\boldsymbol{fp_{1,\ldots k}}$, and the corresponding important score $\boldsymbol{S_{1,\ldots k}}$, where $\boldsymbol{k}$ = 50 for the top50 important fingerprints.

a) Match the important fingerprints $\boldsymbol{fp_{1,\ldots k}}$ on molecule $\boldsymbol{m}$**,** and then average the importance on atoms and bonds of molecule $\boldsymbol{m}$**:**

atoms importance list $\boldsymbol{atoms_m} = []$

bonds importance list $\boldsymbol{bonds_m} = []$

for each $\boldsymbol{fp_i}$ in $\boldsymbol{fp_{1,\ldots k}}$, do**:**

if $\boldsymbol{m}$ contains fingerprint $\boldsymbol{fp_i}$: *# match the substructure or pharmacophore triplet pattern*

for each atom $\boldsymbol{a}$ in molecule $\boldsymbol{m}$ that is matched by $\boldsymbol{fp_i}$:

$\boldsymbol{atoms_m}$ append ($\boldsymbol{a}$, $\boldsymbol{S_i}$)

for each bond $\boldsymbol{b}$ in molecule $\boldsymbol{m}$ that is matched by $\boldsymbol{fp_i}$:

$\boldsymbol{bonds_m}$ append ($\boldsymbol{b}$, $\boldsymbol{S_i}$)

Finally, group by $\boldsymbol{atoms_m}$ / $\boldsymbol{bonds_m}$ on atoms/ bonds and apply mean operation, to get the each atom/bond average importance in the molecule $\boldsymbol{m}$**.**

b) Highlight molecule $\boldsymbol{m}$ based on the average atoms and bonds importance

References

1 Duvenaud, D. K. *et al.* Convolutional networks on graphs for learning molecular fingerprints. in *Advances in Neural Information Processing Systems 28.* 2224-2232 (NeurIPS, 2015).

2 Xiong, Z. *et al.* Pushing the Boundaries of Molecular Representation for Drug Discovery with the Graph Attention Mechanism. *J. Med. Chem.* **16**, 8749–8760 (2019).

3 Yang, K. *et al.* Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370-3388 (2019).

4 Goh, G. B., Hodas, N. O., Siegel, C. & Vishnu, A. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties. *arXiv preprint arXiv:1712.02034* (2017).

5 Winter, R., Montanari, F., Noé, F. & Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **10**, 1692-1701 (2019).

6 Goh, G. B., Siegel, C., Vishnu, A., Hodas, N. O. & Baker, N. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. *arXiv preprint arXiv:1706.06689* (2017).

7 Goh, G. B., Siegel, C., Vishnu, A. & Hodas, N. O. Chemnet: A transferable and generalizable deep neural network for small-molecule property prediction. *arXiv preprint arXiv:1712.02734* (2017).

8 McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:.03426* (2018).

9 Cortés-Ciriano, I. & Bender, A. KekuleScope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images. *J. Cheminformatics* **11**, 41 (2019).

10 Hall, L. H. & Kier, L. B. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **35**, 1039-1045 (1995).

11 Nikolaus Stiefl, I. A. W., Knut Baumann, Andrea Zaliani. ErG: 2D pharmacophore descriptions for scaffold hopping. *J. Chem. Inf. Model.* **46**, 208-220 (2006).

12 McGregor, M. J. & Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Model.* **39**, 569-574 (1999).

13 Gedeck, P., Rohde, B. & Bartels, C. QSAR− How good is it in practice? Comparison of descriptor sets on an unbiased cross section of corporate data sets. *J. Chem. Inf. Model.* **46**, 1924-1936 (2006).

14 Landrum, G. RDKit Documentation Release 2019.09.1. (2019).

15 Probst, D. & Reymond, J.-L. A probabilistic molecular fingerprint for big data settings. *J. Cheminformatics* **10**, 66 (2018).

16 Capecchi, A., Probst, D. & Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminformatics* **12**, 1-15 (2020).

17 Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513-530 (2018).

18 Li, X., Xu, Y., Lai, L. & Pei, J. Prediction of human cytochrome P450 inhibition using a multitask deep autoencoder neural network. *Mol. Pharm.* **15**, 4336-4345 (2018).

19 Wenzel, J., Matter, H. & Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J. Chem. Inf. Model.* **59**, 1253-1268 (2019).

20 Mayr, A. *et al.* Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* **9**, 5441-5451 (2018).

21 Ramsundar, B., Eastman, P., Walters, P. & Pande, V. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*. (" O'Reilly Media, Inc.", 2019).

22 Bento, A. P. *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **42**, D1083-D1090 (2014).

23 Mullard, A. BACE inhibitor bust in Alzheimer trial. *Nat. Rev. Drug Discov.* **16** (2017).

24 Bongarzone, S. & Gee, A. D. BACE1: Now we can see you. *J. Med. Chem.* **61**, 3293–3295 (2018).

25 Probst, D. & Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminformatics* **12**, 1-13 (2020).

26 Altmann, A., Toloşi, L., Sander, O. & Lengauer, T. Permutation importance: a corrected feature importance measure. *Bioinformatics* **26**, 1340-1347 (2010).

27 Fisher, A., Rudin, C. & Dominici, F. Model class reliance: Variable importance measures for any machine learning model class, from the “Rashomon” perspective. *arXiv preprint arXiv:1801.01489* **68** (2018).