nature machine intelligence

ARTICLES

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

Check for updates

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

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] ✉

Successful deep learning critically depends on the representation of the learned objects. Recent state-of-the-art pharmaceutical deep learning models successfully exploit graph-based de novo learning of molecular representations. Nonetheless, the combined potential of human expert knowledge of molecular representations and convolution neural networks has not been adequately explored for enhanced learning of pharmaceutical properties. Here we show that broader exploration of human-knowledge-based molecular representations enables more enhanced deep learning of pharmaceutical properties. By broad learning of 1,456 molecular descriptors and 16,204 fingerprint features of 8,506,205 molecules, a new feature-generation method MolMap was developed for mapping these molecular descriptors and fingerprint features into robust two-dimensional feature maps. Convolution-neural-network-based MolMapNet models were constructed for out-of-the-box deep learning of pharmaceutical properties, which outperformed the graph-based and other established models on most of the 26 pharmaceutically relevant benchmark datasets and a novel dataset. The MolMapNet learned important features that are consistent with the literature-reported molecular features.

The quality and efficiency of deep learning critically depends on the representation of the learned objects. In particular, enhanced pharmaceutical learning depends on appropriate molecular representations (MolRs)[1,2]. By learning their own optimized representations directly from the underlying graphs of the molecules, graph-based de novo learning of MolRs has enabled notably improved deep learning of pharmaceutical and physicochemical properties, outperforming those based on conventionally used molecular descriptors (MolDs) and fingerprint features (FFs)[3–5]. The graph-based approaches may in some cases be subject to limited information processing across the graphs[6]. Therefore, broader exploration of MolRs complements graph-based and other methods for more enhanced deep learning of pharmaceutical properties.

Many MolDs and FFs have been derived from human expert knowledge for comprehensive presentation of the constitutional, physicochemical, topological, structural and substructural features of molecules[7,8], which are valuable priors for feature generation and deep learning of pharmaceutical properties. But questions remain regarding how these priors can be featurized into more appropriate representations. In general, good representations are task non-specific priors that capture posterior distribution of the multiple underlying explanatory elements, enable disentangling and clustering of these elements, and support smooth and flexible local generalization of task functions[9]. For instance, the destruction–construction learning method is capable of recognizing highly difficult fine-grained images[10]. In destruction–construction learning, input images are partitioned into local regions, which are shuffled for exposing discriminative local features and then reconstructed for revealing semantic cross-region correlation relationships, leading to state-of-the-art (SOTA) performance on three benchmark datasets[10].

Therefore, appropriate feature generation may be established by broad profiling of the intrinsic correlations of diverse sets of MolDs and FFs with respect to a large number of molecules in the known chemical space. Moreover, converting one-dimensional unordered vectors to two-dimensional (2D) clustered feature maps (Fmaps) enable efficient learning (parameters saving) using shared-weights architectures of convolution neural networks (CNNs)[11,12]. The development of such feature-generation methods may be facilitated by the extensive studies of MolRs, open-source tools[7,8,13], correlation metrics[14–16], quantification of chemical diversity and characteristics[17,18], and dimensionality reduction methods[19,20].

In this Article, we have developed a new molecular feature-generation method MolMap for mapping MolDs and FFs into robust 2D Fmaps that capture the intrinsic correlations of molecular features. MolMap was trained by broadly profiling 1,456 MolDs and 16,204 FFs of 8,506,205 molecules. MolMap representations were evaluated for out-of-the-box (OOTB) deep learning of 13 pharmaceutical and 3 physicochemical properties on 26 public benchmark datasets. A CNN architecture MolMapNet was constructed as an OOTB development tool for automated deep learning applications, wherein the same set of default parameters were set up for all learning tasks. OOTB tools aim at taking human out of the learning processes, allowing more people to use them[21]. The performances of the MolMapNet models were evaluated with respect to those of the SOTA deep learning models on the same benchmark datasets and

[1]The State Key Laboratory of Chemical Oncogenomics, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, P. R. China. [2]Bioinformatics and Drug Design Group, Department of Pharmacy, and Center for Computational Science and Engineering, National University of Singapore, Singapore, Singapore. [3]Department of Biological Medicines and Shanghai Engineering Research Center of Immunotherapeutics, Fudan University School of Pharmacy, Shanghai, P. R. China. [4]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, P. R. China. [5]Shenzhen Kivita Innovative Drug Discovery Institute, Shenzhen, P. R. China. ✉e-mail: jiangyy@sz.tsinghua.edu.cn; phacyz@nus.edu.sg

Table 1 | Summary of benchmark datasets in this study

Data class	Dataset and split reference	Number of molecules	Number of tasks	Task metric	Task type
Physicochemical	ESOL (estimating the aqueous solubility), water solubility[3,4]	1,128	1	RMSE	Regression
	FreeSolv, solvation free energy[3,4]	642	1	RMSE	Regression
	Lipop, lipophilicity[3,4]	4,200	1	RMSE	Regression
Molecular binding	PDBbind-F, PDBbind-C and PDBbind-R, ligand–protein binding full, core and refined (three datasets)[3,4]	9,880, 168, 3,040	1 for each	RMSE	Regression
Bioactivity	PCBA, PubChem HTS Bioassay[3]	437,929	128	PRC-AUC	Classification
	MUV (maximum unbiased validation) PubChem Bioassay[3]	93,087	17	PRC-AUC	Classification
	ChEMBL bioassay activity dataset[3,69]	456,331	1,310	ROC-AUC	Classification
	Cancer cell-line IC50 A2780, CCRF-CEM12, DU-14512, HCT-1512, KB12, LoVo12, PC-312 and SK-OV-312 (eight datasets)[27]	2,255, 3,047, 2,512, 994, 2,731, 1,120 4,294, 1,589	1 for each	R^2	Regression
	Malaria, anti-malarial EC50[4]	9,998	1	RMSE	Regression
	BACE (beta-secretase 1) inhibitors[3,4]	1,513	1	ROC-AUC	Classification
	HIV (human immunodeficiency virus) replication inhibition[3,4]	41,127	1	ROC-AUC	Classification
Toxicity	Tox21, toxicology in the twenty-first century[3,4]	7,831	12	ROC-AUC	Classification
	SIDER (side effect resource), adverse drug reactions of marketed drugs[3,4]	1,427	27	ROC-AUC	Classification
	ClinTox, clinical trial toxicity[3,4]	1478	2	ROC-AUC	Classification
Pharmacokinetic	CYP (cytochrome P450) PubChem Bioassay CYP 1A2, 2C9, 2C19, 2D6, 3A4 inhibition[32]	16,896	5	ROC-AUC	Classification
	LMC-H, LMC-R and LMC-M, liver microsomal clearance in human, rat and mouse[28]	8,755	3	R^2	Regression
	BBBP, blood–brain barrier penetration[3,4]	2,039	1	ROC-AUC	Classification

Each dataset was split into training, validation and test sets by using the corresponding data-split code of published studies (reference provided).

data splits. MolMap and MolMapNet open-source libraries are at https://github.com/shenwanxiang/bidd-molmap.

Background

Deep learning of pharmaceutical properties has been conducted based on four MolR classes (Supplementary Fig. 1 and Supplementary Table 1). The first is graph-based feature representations, where graph convolutional networks (GCNs) or graph attention networks (GATs) have been explored for de novo learning directly from the underlying graphs of molecules[4–6], leading to the SOTA performances on pharmaceutically related tasks[22]. The second is string-based representations, where CNN and recurrent neural networks have been employed for learning from the embeddings of the string representations of chemical structures (for example, canonical simplified molecular-input line-entry system (SMILES))[23–25]. The third is the image representations, where CNNs have been used for learning from the rule-based renderings of a 2D chemical digital grid or Kekulé images[26,27]. The fourth is knowledge-based representations, where deep learning models have been developed for learning from the priori human-knowledge-derived MolDs or FFs[28].

Although it is preferable to explore the lower-level representations without relying on human intuitions, the extensive knowledge bases of MolDs and FFs are highly useful for learning MolRs and pharmaceutical properties from human-knowledge perspectives. In particular, subsets of MolDs and FFs show a high degree of correlation, which provides unique clues for appropriate MolRs. Some MolDs or FFs are related by design (for example, MolWeight and MolExactWeight), while some 'unrelated' ones show high degrees of intrinsic correlation. Investigations of chemical screening collections have revealed that polar surface area correlates with the counts of hydrogen bond acceptors and donors[17]. The clustering of these correlated MolDs or FFs and their projection into 2D Fmaps enable feature pattern agglomeration for efficient learning by the shared-weights CNN architectures[29].

For coordinated learning of MolDs and FFs, it is desirable to use a universal correlation metric for both MolDs and FFs. Cosine correlation has consistently performed comparably well as the widely used Tanimoto coefficient in certain FF-based molecular studies[15,30] and the widely used Euclidean distance in some MolD-based classification tasks[16]. Therefore, cosine correlation may be used for MolD/FF-based feature generation. To learn from MolDs and FFs with CNNs, high-dimensional MolDs and FFs need to be projected into 2D Fmaps, which requires a manifold learning algorithm with minimal loss of information. The recently developed uniform manifold approximation and projection (UMAP) tool[20], based on the Riemannian geometry and algebraic topology algorithms, has demonstrated competitive capability for this task[19].

Results and discussion

MolMap Fmaps. Using the MolMap package (Fig. 1), we generated the MolD and FF Fmaps of aspirin and its analogue *N*-acetylanthranilic acid (Fig. 2). Although these molecules are highly similar in structure, their MolD Fmaps contain small areas of markedly different patterns and their FF Fmaps contain regions of substantially different patterns. These patterns (for example, the purple and light-blue dashed boxes of Fig. 2) are capturable by a

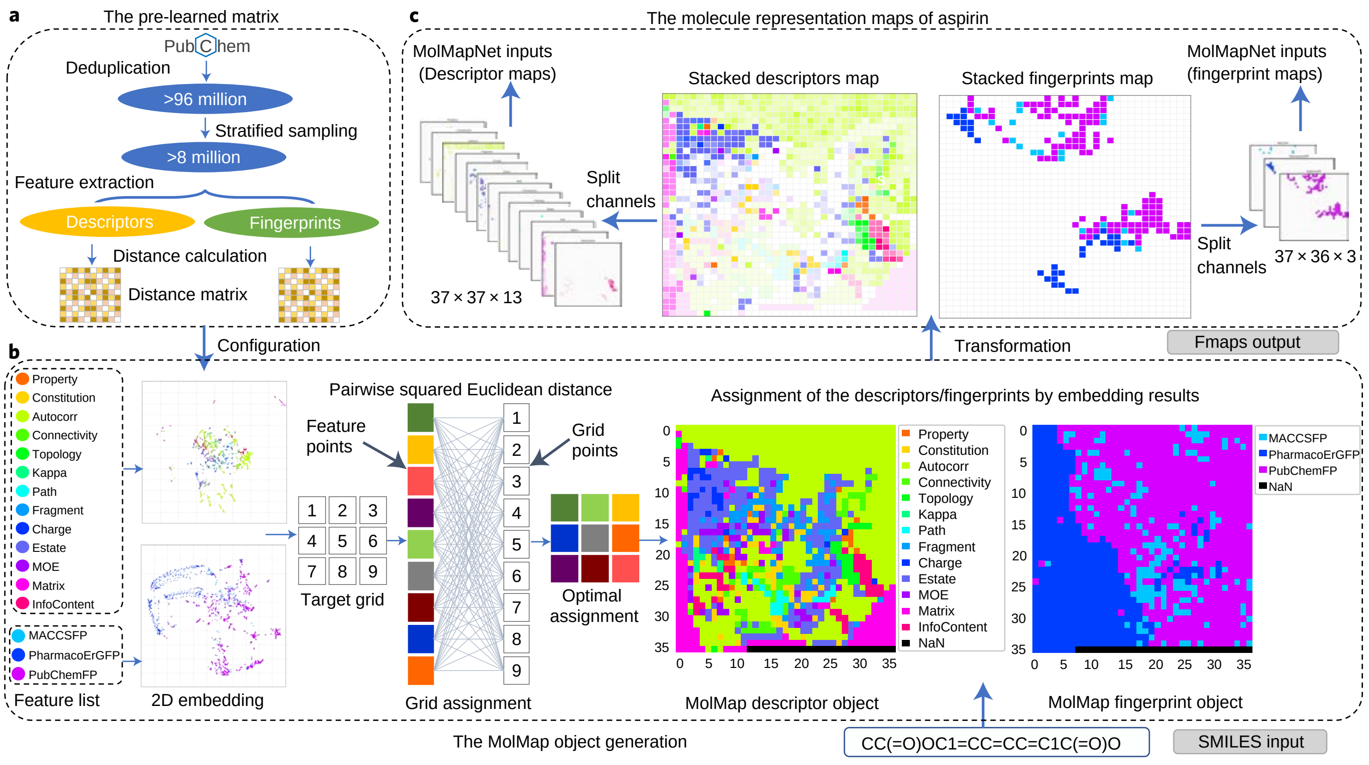


Fig. 1 | MolMap feature-generation flowchart. a, Derivation of the pre-trained distance matrix of molecular descriptors and fingerprints by broad profiling of PubChem molecules. **b**, Extraction of MolMap distance matrix of user-selected molecular descriptors and/or fingerprints from the pre-trained distance matrix, followed by UMAP projection of these descriptors and fingerprints into the respective 2D embedding and the subsequent mapping into the MolMap 2D Fmap by using the J–V algorithm. **c**, Generation of the MolMap Fmaps of 13 classes of descriptors and 3 sets of fingerprints for aspirin, upon inputting its SMILES string in **b** (bottom right).

typical CNN filter. In the MolD Fmaps, different MolD classes are primarily concentrated in distinctive areas. In the FF Fmaps, the PharmacoErGFP FFs are largely separated from the other FFs, and the MACCSFP and PubChemFP FFs are complementarily located in the same regions. Moreover, the correlated MolDs or FFs form clusters. For instance, three quantitative estimate of druglikeness (QED)[18] MolDs (MolQedWeightsMax, MolQedWeightsMean and MolQedWeightsNone) are clustered together (Fig. 2) and various other correlated MolDs are also clustered together (Supplementary Fig. 2). These indicate that MolMap Fmaps present distinguished representations and intrinsic correlations of molecular and structural features.

MolMapNet deep learning performances with respect to the SOTA graph-based GCN/GAT models. The GCNs/GATs have achieved SOTA performances on a number of benchmark datasets[3–5]. Among these GCN-/GAT-tested datasets, there are 13 pharmaceutical (3 molecular binding, 6 bioactivity, 3 toxicity, 1 pharmacokinetic) and 3 physicochemical datasets with available data-split codes. Therefore, MolMapNet OOTB models (Fig. 3) were developed on these 16 datasets and compared with the published performances of the GCN/GAT models (Table 2) using the same data split, evaluation metric and (for multitask datasets) multitask training method[3–5] (Supplementary Method 1). MolMapNet outperformed the MoleculeNet models[3] and directed message passing neural networks (D-MPNN) models[5] on 9 of the 12 pharmaceutical datasets but underperformed these GCN models on all 3 physicochemical datasets (Table 2). MolMapNet also outperformed the AttentiveFP models[4] on 7 of the 8 pharmaceutical datasets but underperformed these GAT models on all 3 physicochemical datasets. These results suggested that MolMap Fmaps are highly appropriate MolRs, and MolMapNet is useful for learning pharmaceutical properties competitively with respect to the SOTA. MolMapNet underperformed the GCN/GAT models on the physicochemical datasets partly for the following reason: MolMapNet learns from MolDs, some of which are computed physicochemical properties (for example, the calculated logP, clogP). MolMapNet learning of physicochemical properties is subject to the intrinsic errors in the computed MolDs of physicochemical property values (for example, clogP values of drugs slightly differ from experimental values by a correlation coefficient 0.955 (ref. [31])). GCN/GAT de novo learning algorithms avoid these intrinsic errors and thus are more advantageous for learning physicochemical properties.

The performance of the MolMapNet OOTB models was further evaluated over 10 different random seeds of data splits with respect to those of the D-MPNN models[5] and AttentiveFP models[4] on 12 benchmark datasets (Extended Data Figs. 1 and 2). Except the physicochemical property prediction tasks, MolMapNet consistently showed better performance with respect to different seeds, and the performance was at comparable or smaller variations than those of the D-MPNN and AttentiveFP models. For the three physicochemical property prediction tasks, MolMapNet was mostly outperformed by the D-MPNN and AttentiveFP models but nonetheless exhibited similar patterns of variations as one or both D-MPNN and AttentiveFP models.

MolMapNet deep learning performances with respect to the chemical graph-based CNN models. The chemical graph-based CNN models have performed well in deep learning of pharmaceutical and physicochemical properties without the chemical knowledge[26,27]. Some of these CNN models are as deep as 19 layers[27] for end-to-end learning, while MolMapNet is a CNN of fewer layers.

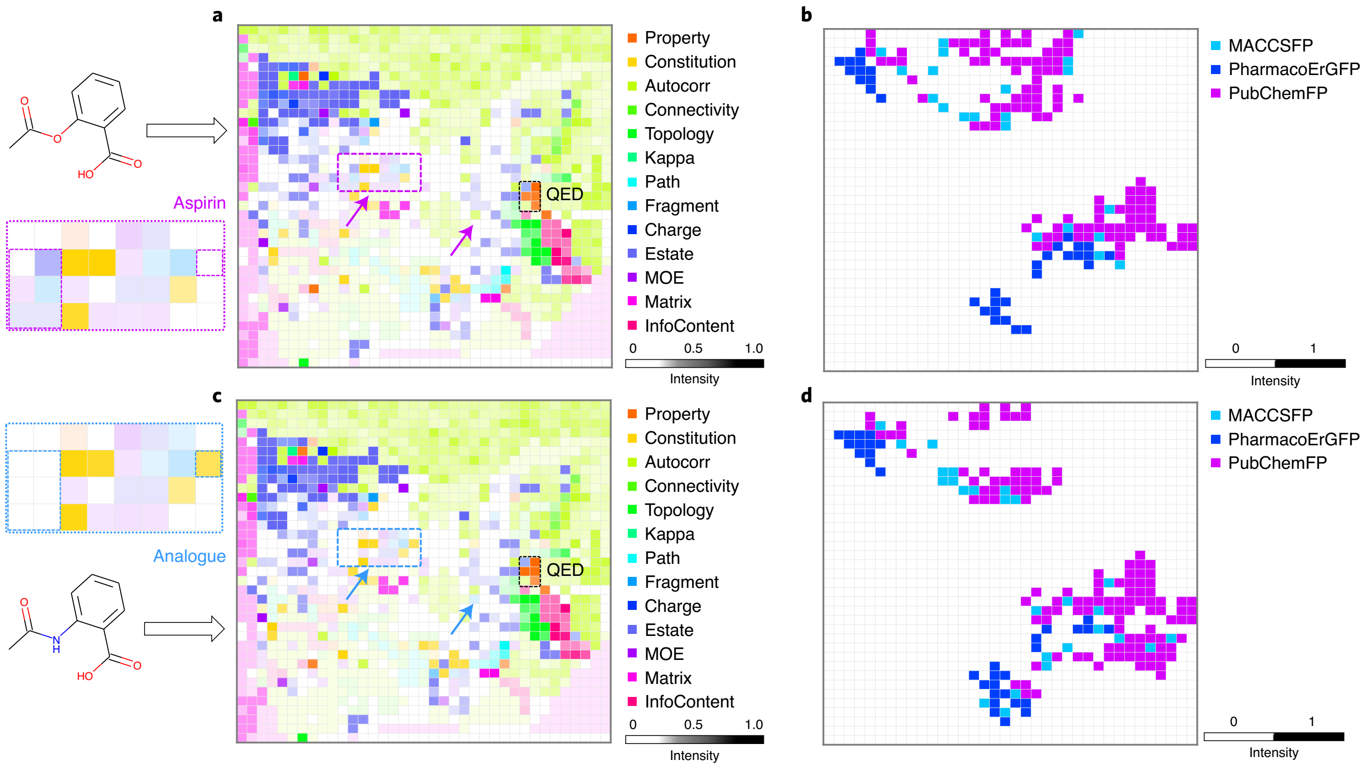


Fig. 2 | MolMap multichannel descriptor and fingerprint Fmaps. a–d, MolMap Fmaps of 13 classes of descriptors (**a**,**c**) and 3 sets of fingerprints (**b**,**d**) of aspirin (**a**,**b**) and its analogue *N*-acetylanthranilic acid (**c**,**d**). Each descriptor class or fingerprint set is in a distinctive colour (channel) and the corresponding feature values are indicated by the colour intensities (higher values are with higher intensities, 0 value is in white). The three clustered QED descriptors are presented as the red/orange dots in the black dashed rectangular box. The purple arrow in **a** and the light-blue arrow in **c** indicate the regions of differential patterns between the MolMap Fmap of aspirin and its analogue. These patterns are highlighted in the purple and light-blue dashed rectangular boxes between the aspirin and analogue structures on the left. The grey-scale colour bar stands for the intensity of the feature point values for different subtypes.

Differential performances of MolMapNet with respect to these CNN models partly reflect the differential capacity of MolMap Fmaps for learning pharmaceutical properties. Notably, the image-based 19-layer CNN KekuleScope models have recorded outstanding performances on 8 cancer cell-line benchmark datasets[27], the performance of these models may be compared to MolMapNet models because the datasets and data-split codes of these models are available. Hence, MolMapNet OOTB models were developed on these 8 benchmark datasets and compared with the published performances of the KekuleScope models[27] (Table 3) using the same data split and evaluation metric[27] (Supplementary Method 1). MolMapNet outperformed the KekuleScope models on all 8 datasets (squared Pearson correlation coefficient between predicted and observed values $R^2 = 0.583$–0.734 versus $R^2 = 0.427$–0.622).

MolMapNet multitask deep learning performances with respect to the molecular-descriptor-based multitask fully connected deep neural network models. Investigations have suggested that multitask fully connected deep neural networks (FC-DNNs) can perform better than single-task FC-DNNs in predicting pharmaceutical properties[28,32,33]. In particular, autoencoder (AE)-based[32] and Sanofi-Aventis[28] FC-DNN models have scored outstanding performances on two multitask benchmark datasets: the CYP isoenzyme inhibitor dataset and the liver microsomal clearance dataset. MolMapNet may be competitive in multitask learning by more appropriate MolRs. Thus, multitask MolMapNet OOTB models were developed on these two multitask datasets and compared with the published performances of the AE-based[32] and Sanofi-Aventis[28] models by using the same data split, evaluation metric and multitask training method, respectively (Table 3). On the CYP450 datasets, MolMapNet underperformed the AE-based FC-DNN models[32] on three of the five tasks, but the area under the receiver operating characteristic curve (ROC-AUC) values of the three underperformed tasks are nonetheless comparable to those of the AE-based FC-DNN models. On the liver microsomal clearance tasks, MolMapNet outperformed the Sanofi-Aventis[28] models on all three tasks. Overall, the MolMap Fmaps and the multitask MolMapNet architectures are competitive for multitask learning of pharmaceutical properties.

Single-path versus dual-path MolMapNet deep learning models. Extended Data Fig. 3 shows the comparative performances of the single-path and dual-path MolMapNet OOTB models on 11 benchmark datasets of the MoleculeNet data splits and AttentiveFP data splits. For the regression tasks, the MolD-only single-path (MolMapNet-D) models performed as comparably well as or better than the joint MolD and FF dual-path (MolMapNet-B) models on three of the five regression datasets. For classification tasks, the FF-only single-path (MolMapNet-F) models performed as comparably well as the MolMapNet-B models on four of the six classification datasets. Interestingly, the MolMapNet-F models performed slightly worse in regression tasks but slightly better in classification tasks than the MolMapNet-D models. Interestingly, the input Fmaps of MolMapNet-D models are quantitative MolDs (for example, molecular weight), while the input Fmaps of MolMapNet-F models are categorical FFs (0 or 1) (Fig. 2). Consequently, MolMapNet-D

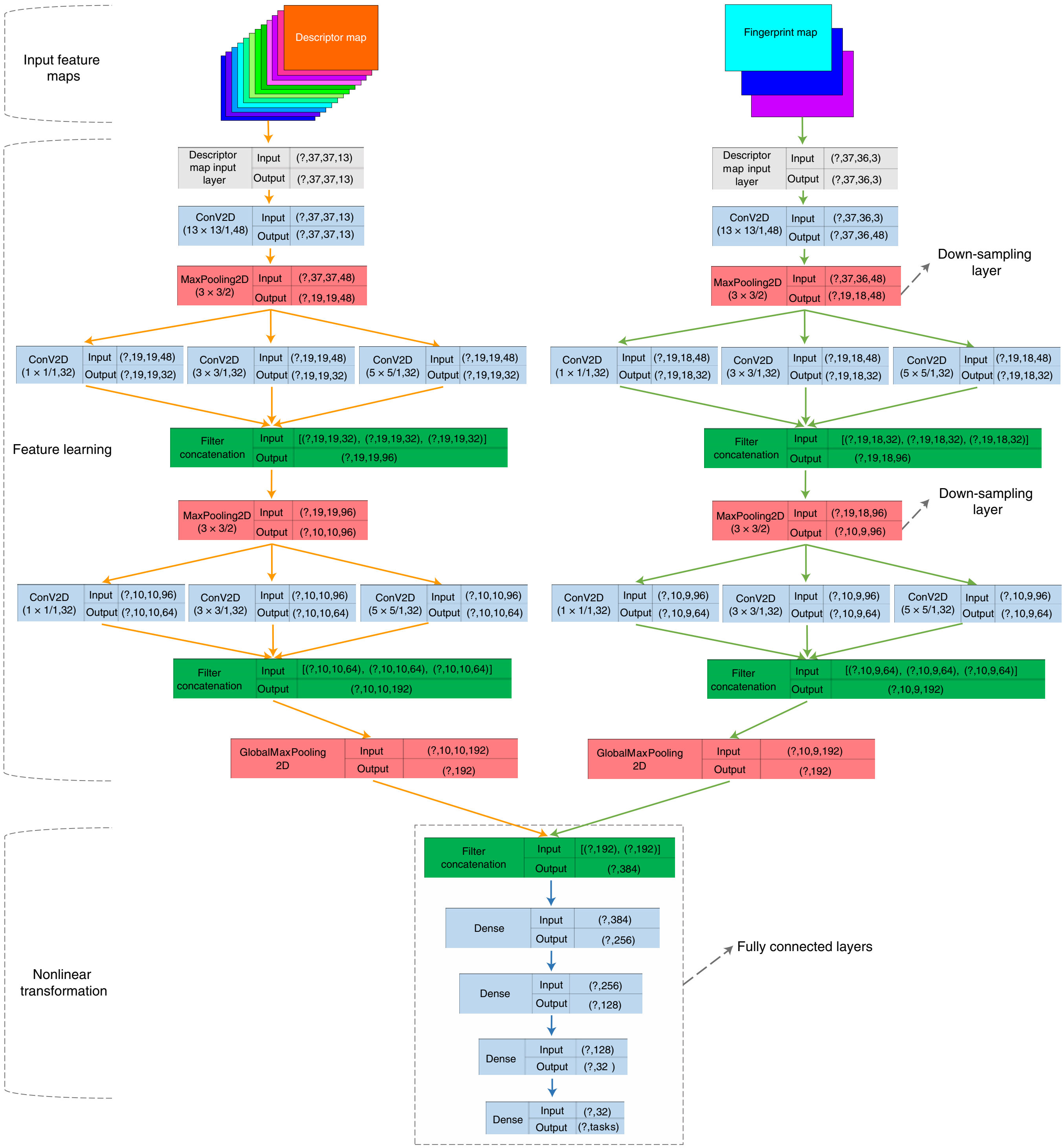

Input feature maps
Descriptor map
Fingerprint map
Descriptor map input layer | Input (?,37,37,13) | Output (?,37,37,13)
Descriptor map input layer | Input (?,37,36,3) | Output (?,37,36,3)
ConV2D (13 × 13/1,48) | Input (?,37,37,13) | Output (?,37,37,13)
ConV2D (13 × 13/1,48) | Input (?,37,36,3) | Output (?,37,36,48)
MaxPooling2D (3 × 3/2) | Input (?,37,37,48) | Output (?,19,19,48)
MaxPooling2D (3 × 3/2) | Input (?,37,36,48) | Output (?,19,18,48)
Down-sampling layer
Feature learning
ConV2D (1 × 1/1,32) | Input (?,19,19,48) | Output (?,19,19,32)
ConV2D (3 × 3/1,32) | Input (?,19,19,48) | Output (?,19,19,32)
ConV2D (5 × 5/1,32) | Input (?,19,19,48) | Output (?,19,19,32)
ConV2D (1 × 1/1,32) | Input (?,19,18,48) | Output (?,19,18,32)
ConV2D (3 × 3/1,32) | Input (?,19,18,48) | Output (?,19,18,32)
ConV2D (5 × 5/1,32) | Input (?,19,18,48) | Output (?,19,18,32)
Filter concatenation | Input [(?,19,19,32), (?,19,19,32), (?,19,19,32)] | Output (?,19,19,96)
Filter concatenation | Input [(?,19,18,32), (?,19,18,32), (?,19,18,32)] | Output (?,19,18,96)
MaxPooling2D (3 × 3/2) | Input (?,19,19,96) | Output (?,10,10,96)
MaxPooling2D (3 × 3/2) | Input (?,19,18,96) | Output (?,10,9,96)
Down-sampling layer
ConV2D (1 × 1/1,32) | Input (?,10,10,96) | Output (?,10,10,64)
ConV2D (3 × 3/1,32) | Input (?,10,10,96) | Output (?,10,10,64)
ConV2D (5 × 5/1,32) | Input (?,10,10,96) | Output (?,10,10,64)
ConV2D (1 × 1/1,32) | Input (?,10,9,96) | Output (?,10,9,64)
ConV2D (3 × 3/1,32) | Input (?,10,9,96) | Output (?,10,9,64)
ConV2D (5 × 5/1,32) | Input (?,10,9,96) | Output (?,10,9,64)
Filter concatenation | Input [(?,10,10,64), (?,10,10,64), (?,10,10,64)] | Output (?,10,10,192)
Filter concatenation | Input [(?,10,9,64), (?,10,9,64), (?,10,9,64)] | Output (?,10,9,192)
GlobalMaxPooling 2D | Input (?,10,10,192) | Output (?,192)
GlobalMaxPooling 2D | Input (?,10,9,192) | Output (?,192)
Nonlinear transformation
Filter concatenation | Input [(?,192), (?,192)] | Output (?,384)
Dense | Input (?,384) | Output (?,256)
Dense | Input (?,256) | Output (?,128)
Dense | Input (?,128) | Output (?,32)
Dense | Input (?,32) | Output (?,tasks)
Fully connected layers

Fig. 3 | MolMapNet deep learning architecture. The architecture comprises three components: multichannel input feature mapping, dual-path CNN feature learning and nonlinear transformation with fully connected layers. The grey boxes are the input layers, the blue boxes are the convolutional layers and dense layers, the red and green boxes are the pooling and concatenation layers, respectively. The numbers in brackets of each box are the sizes of the Fmaps or kernels (For instance, each number in (?, 37, 37, 13) stands for n_{samples}, height, width and n_{channels} for a Fmap, respectively). The molecular descriptors are fed into the left path with a multichannel input layer of up to 13 channels for up to 13 descriptor classes. The fingerprints are fed into the right path with a multichannel input layer of up to 3 channels for up to 3 fingerprint sets. Each path is dedicated to the feature learning of either descriptors or fingerprints. The dual-path latent features are concatenated and subsequently fed to the fully connected layers for nonlinear transformation and task execution. Trainable parameters: left path, ~0.40 million; right path, ~0.32 million; dual path, ~0.80 million.

models tend to perform better for continuous predictors such as regression tasks, while MolMapNet-F models are better for categorical predictors such as classification tasks. Overall, MolMapNet-B models take advantage of both input types (MolD and FF Fmaps), thereby becoming highly completive in both regression and classification tasks.

Table 2 | MolMapNet performances on 15 benchmark datasets compared with the graph-based models

Data class	Dataset	Task metric	MoleculeNet[3]	Chemprop[5]	AttentiveFP[4]	MMNB (OOTB)
			(GCN best)	(D-MPNN)		
Physicochemical	ESOL	RMSE	0.580 (MPNN)	**0.555**		0.575
					0.486	0.543
	FreeSolv	RMSE	1.150 (MPNN)	**1.075**		1.155
					0.773	0.994
	Lipop	RMSE	0.655 (GC)	**0.555**		0.625
					0.564	0.640
Molecular binding	PDBbind-F	RMSE	1.440 (GC)	1.391		**0.721**
					0.766	**0.753**
	PDBbind-C	RMSE	1.920 (GC)	2.173		**0.931**
	PDBbind-R	RMSE	1.650 (GC)	1.486		**0.889**
Bioactivity	Malaria	RMSE			1.077	**1.011**
	BACE	ROC-AUC	0.806 (Weave)	N/A		**0.849**
					0.856	**0.881**
	HIV	ROC-AUC	0.763 (GC)	0.776		**0.777**
					0.848	**0.865**
	MUV	PRC-AUC	**0.109** (Weave)	0.041		0.096
	PCBA	PRC-AUC	0.136 (GC)	**0.335**		0.276
	ChEMBL	ROC-AUC		0.739		**0.750**
Toxicity	Tox21	ROC-AUC	0.829 (GC)	0.851		0.845
					0.845	0.842
	SIDER	ROC-AUC	0.638 (GC)	0.676		**0.680**
					0.640	**0.700**
	ClinTox	ROC-AUC	0.832 (GC)	0.864		**0.888**
					0.945	**0.973**
Pharmacokinetic	BBBP	ROC-AUC	0.690 (Weave)	0.738		**0.739**
					0.931	**0.961**

The MolMapNet, Chemprop and the AttentiveFP models use the same dataset, data split and (for multitask datasets) multitask training method as the MoleculeNet models or the AttentiveFP models, respectively. The bold indicates the cases of the models outperforming all other models. MPNN, message passing neural networks; GC, graph convolutional models; Weave, Weave models. MMNB-OOTB: out-of-the-box performance of MolMapNet. N/A, not avaliable.

MolMapNet deep learning performances with respect to the *k*-nearest-neighbour models. The performances of the MolMapNet OOTB models are partly attributable to the pre-training of the MolMap Fmaps. To probe the influence of this pre-training on classification performances, we evaluated the models with and without the pre-training of the FF Fmaps. Specifically, the performance of the MolMapNet-F OOTB models was compared with the *k*-nearest neighbour (kNN) models (Supplementary Method 2), which were built from the same three FF sets as the MolMapNet-F models on five classification benchmark datasets (BACE, BBBP, HIV, ClinTox and SIDER) and the MoleculeNet data splits (Extended Data Fig. 4). MolMapNet-F OOTB models outperformed the kNN models for four of the five datasets by noticeable margins, that is, the ROC-AUC values are 0.843 versus 0.851, 0.744 versus 0.677, 0.774 versus 0.728, 0.869 versus 0.806, and 0.684 versus 0.630 for the BACE, BBBP, HIV, ClinTox and SIDER datasets, respectively. In contrast, the ROC-AUC values of the best of the MoleculeNet[3] and Chemprop[5] models are 0.806, 0.738, 0.776, 0.864 and 0.676 for the BBBP, ClinTox, HIV and SIDER datasets, respectively (Table 2). Therefore, MolMap pre-training is advantageous in enhanced learning of pharmaceutical properties. Noticeably, all kNN models performed well, with the BACE model outperforming the GCN and MolMapNet OOTB models. These performances are partly attributable to the appropriate MolRs by the three MolMap-selected FF sets. These FF sets were selected because their MolMap Fmaps present distinguished and more densely clustered patterns than the other FF sets, thereby facilitating enhanced learning (Methods).

Optimized versus OOTB MolMapNet deep learning models. We optimized four hyperparameters of the MolMapNet-B models by coarse-grained grid search: the UMAP feature-generation parameters for enhanced feature representation, the batch size for improved convergence and performance, the kernel size of the first convolution layer for more effective receptive field, and the dense layer width for improved multitask classification. First, each hyperparameter was individually optimized. The UMAP optimization on the ESOL, BACE and Tox21 datasets showed that a reduced number of neighbours usually boosts performance by increased precision of the local feature distribution[19] (Extended Data Fig. 5). The batch size optimization on the ESOL and FreeSolv datasets revealed that decreased batch sizes accelerates convergence and improves performance (Supplementary Fig. 3). The kernel size optimization on the BACE dataset suggested that increased kernel size enhances learning via a more effective receptive field[34] (Supplementary Fig. 4). The dense layer width optimization indicates that increased number of nodes improves the multitask performances by expanded information

Table 3 | MolMapNet performance on eight single-task and two multitask benchmark datasets compared with CNN and fully connected DNN models

Dataset	Individual task	Task metric	Task performance	
Single task[a]			**KekuleScope[27] (VGG19-bn)**	**MolMapNet (MMNB)**
Cancer cell-line bioactivity	A2780	R^2	0.622	**0.663**
	CCRF-CEM	R^2	0.528	**0.627**
	DU-145	R^2	0.427	**0.594**
	HCT-15	R^2	0.617	**0.734**
	KB	R^2	0.533	**0.713**
	LoVo	R^2	0.530	**0.583**
	PC-3	R^2	0.496	**0.615**
	SK-OV-3	R^2	0.461	**0.597**
Multitask joint training method[b]			**AE-based FC-DNN[32]**	**MolMapNet (MMNB)**
CYP isoenzyme inhibitors	1A2	ROC_AUC	**0.982**	0.975
	2C9	ROC_AUC	0.799	**0.805**
	2D6	ROC_AUC	0.878	**0.908**
	2C19	ROC_AUC	**0.832**	0.823
	3A4	ROC_AUC	**0.929**	0.923
Multitask alternate training method[b]			**Sanofi-Aventis FC-DNN[28]**	**MolMapNet (MMNB)**
Liver microsomal clearance	LMC-H	R^2	0.566	**0.580**
	LMC-R	R^2	0. 771	**0.790**
	LMC-M	R^2	0.475	**0.526**

[a]Eight single-task cancer cell-line bioactivity benchmark datasets compared with the CNN-based KekuleScope models using the same dataset and the same data split. [b]Two multitask benchmark datasets compared with the AE-based and Sanofi-Aventis FC-DNN models using the same dataset, data split and the multitask training method. VGG19-bn, VGG19 with batch normalization; Sanofi-Aventis FC-DNN, multitask fully connected deep neural networks developed by a Sanofi-Aventis team. The bold indicates the cases of the model outperforming the comparative model.

processing capacity beyond the OOTB settings tailored mostly to single task. Second, the four hyperparameters were collectively optimized on eight datasets (ESOL, FreeSolv, Malaria, BACE, HIV, MUV, PCBA, ChEMBL) using the MoleculeNet and AttentiveFP data splits (Supplementary Table 5). Before optimization, only 5 of the 12 MolMapNet-B models outperformed the SOTA GCN/GAT models. After optimization, nine models outperformed the SOTA GCN/GAT models under the same data splits. In particular, the RMSE of the FreeSolv model is reduced by 14.7% (from 1.075 to 0.916) and the area under the precision recall curve (PRC-AUC) value of the MUV model is increased by 44.9% (from 0.109 to 0.158). For the three underperforming MolMapNet-B models, their MolMapNet-D counterparts are substantially better, with two models outperforming the SOTA AttentiveFP GAT models (RMSE 0.477 versus 0.486 on the ESOL dataset and 0.728 versus 0.773 on the FreeSolv dataset) in the same data splits (Extended Data Fig. 3).

MolMapNet generalization capability on novel compounds. MolMapNet was evaluated on 216 and 179 novel BACE high-potency inhibitors and low-potency inhibitors (NBACE dataset, Supplementary Table 6) extracted from the ChEMBL database[35] (Supplementary Method 3). The molecular similarity patterns between the NBACE and BACE datasets were visualized by TMAP[36], the Tanimoto coefficients of the compounds between the NBACE and BACE datasets are 0.23–0.61 (Extended Data Fig. 6). Tanimoto coefficients <0.7 typically indicate remote similarity[37]. Thus, the NBACE dataset is novel with respect to the BACE dataset. The performance of the MolMapNet-F model trained by the BACE dataset was tested by the NBACE dataset in comparison with the D-MPNN[5] and AttentiveFP[4] models. The sensitivity and specificity of the MolMapNet-F model are 70% and 84%, compared with 48% and 81% for the GCN D-MPNN[5] model and 63% and 63% for the AttentiveFP[4] model, respectively.

MolMapNet learned deep latent features and important input features. To probe the MolMapNet learned deep latent features and important input features, we first analysed the MolMapNet-D solubility model trained on the ESOL dataset and the AttentiveFP data split. Principal component analysis (PCA) analysis of the latent features of the global max-pooling layer (before the fully connected layers) indicated that these latent features are clustered according to solubility values (Extended Data Fig. 7a). Therefore, task-oriented clustering is a characteristic of deep latent features. The important input features were derived based on an importance score computed from the permutation algorithm[38] and the mean squared error (MSE) metric (Supplementary Method 4). The important input features derived from the training and test sets are correlated (Pearson $r=0.92$, Extended Data Fig. 8a). The top important input features E-state, QED, charge and topological index (Supplementary Table 7) are clustered together (Extended Data Fig. 8b). The E-state index encodes topological environment and electronic interactions relevant for solubility prediction[39–42]. QED descriptors quantify drug-likeness and indicate solubility and permeability of oral drugs[18]. The charge descriptor reflects ionic interactions that affect solubility[43]. We next analysed the important input features of the MolMapNet-F BACE inhibitor model trained on the BACE dataset. The top-ranked important FFs include a group of five PubChemFP FFs (Extended Data Fig. 9a) and a group of seven MACCSFP and PubChemFP FFs (Extended Data Fig. 9b), which are frequently found in BACE high-potency inhibitors but less so in the low-potency inhibitors (Extended Data Fig. 9c).

The top-50 FFs in importance scores were mapped to the individual atoms and bonds of each molecule (Supplementary Methods), which can be visually displayed using different colour schemes to reveal the substructures deemed by deep learning models as important. When analysing the typical 2-aminoquinoline inhibitors[44] and their structurally close neighbours the 2-aminobenzimidazole inhibitors[45], the visualization revealed the hydrophobic carbon chain of the high-potency inhibitor BACE_276 as an important substructure for BACE activities, which is consistent with the conclusions from the structure–activity relationship studies[44] (Extended Data Fig. 10). The top-50 important features may be exploited for indicating potent BACE inhibitory scaffolds (Supplementary Fig. 5). By this approach, 25 of 26 collected clinical trial BACE inhibitor drugs were identified as high-potency inhibitors, while the remaining drug PF-04976081 lacks an identifiable highly important substructure partly because it is of a novel molecular scaffold (Supplementary Fig. 6). Therefore, our analysis suggested that the MolRs important for the pharmaceutically relevant properties can be well captured by MolMapNet for enhanced prediction of these properties.

Conclusions

Accurate learning and prediction of pharmaceutical properties is a challenging task[46], particularly for low-data cases[47] and novel prediction tasks[48]. Appropriate MolRs are critical for enhanced learning and prediction capabilities[1–3,14,49]. Notable progress has been made in graph-based de novo learning of MolRs[3–6]. On top of these remarkable advances, broader exploration of MolRs helps to add more perspectives for enhanced learning and prediction capabilities. In particular, through broader learning of the extensive priori human-knowledge bases, appropriate MolRs may be derived from the rich reservoir of the constitutional, structural and physicochemical properties in MolDs and the high variety of substructures in FFs, thereby facilitating enhanced deep learning of pharmaceutical properties. New tools such as MolMap facilitate feature generation of MolDs and FFs into 2D Fmaps that capture the intrinsic correlations of molecular features for deep learning applications. On the basis of these Fmaps, the shared-weight CNN architectures can be exploited for enhanced learning and prediction of pharmaceutical properties. To reduce the technical barrier and support wider applications, it is desirable to develop deep learning models as OOTB tools[21]. Therefore, CNN-based deep learning MolMapNet models were developed for OOTB deep learning of pharmaceutical properties, which are highly competitive against established models on most of the 26 benchmark datasets. Deep learning models exploring wider variety of representation and feature-generation strategies (for example, the graph-based DNN fingerprint[6]) have continuously progressed. The collective exploration of these and established strategies enable more enhanced deep learning and prediction of pharmaceutical and other molecular properties.

Methods

Data and processing for MolMap learning. The SMILES codes of 138 million molecule entries were downloaded from PubChem[50] (CID-SMILES.gz). These entries were deduplicated based on their canonical InchI codes (computed by RDkit[8]), leading to 110,913,349 unique molecules. These unique molecules were grouped into 100 classes according to their on-bits counts (NumOnBits) of the ECFP4-like Morgan Fingerprint (MorganFP). The stratified sampling technique was used to extract 8,506,205 sampling molecules from the 100 classes for sampling the 110,913,349 molecules (Supplementary Fig. 7).

Molecular descriptors and fingerprint features. Based on the open-source libraries RDkit[8], Mordred[7], PyBioMed[13] and OpenBabel[51], the MolMap molecular feature module was built for computing 1,456 MolDs and 16,204 FFs. These MolDs include 13 classes of constitutional, physicochemical and topological descriptors (Autocorr, InfoContent, Topology, Path, Connectivity, Kappa, Estate, Charge, Matrix, Fragment, Property, Constitution and MOE (molecular operating environment)) (Supplementary Table 2). The FFs include seven sets of topological path-based features (MorganFP (ECFP-like), AtomPairFP, TorsionFP, RDkitFP[8], AvalonFP[52], MHFP[15] and MAP4[53]), two sets of pharmacophore-based features (PharmacoErGFP[54] and PharmacoPFP[55]) and three sets of substructure-key SMARTS-based features (PubChemFP, MACCSFP and EstateFP[56]), and their default calculation settings are available in Supplementary Table 3.

Distance matrix for molecular descriptors and fingerprint features. Using the cosine correlation function $d\_\text{cosine}_{(x,y)} = 1 - \frac{x \cdot y}{\|x\| \|y\|}$, the pairwise distances among 1,456 MolDs and among 16,204 FFs were computed with respect to 8,506,205 sampling molecules, where x or y is a 8,506,205-dimensional vector, with each component being a MolD or FF of a molecule. These broadly learned pairwise distances were stored in a distance matrix of 1,456 × 1,456 dimensions for the MolDs and 16,204 × 16,204 dimensions for the FFs, respectively. MolMap also provides distance matrices based on the Pearson correlation distance $d\_\text{corr}_{(x,y)} = 1 - \frac{(x-\bar{x})(y-\bar{y})}{\|x-\bar{x}\| \|y-\bar{y}\|}$ and Jaccard distance $d\_\text{jaccard}_{(x,y)} = 1 - \frac{x \cap^{y}}{x \cup^{y}}$ (for FFs only). The variance of some MolDs and FFs is low across the molecules (Supplementary Fig. 8a,b). For low-variance data, optimal filtering is achievable by removing ~15% of the lowest variance descriptors[57]. We found that the removal of ~10% of the lowest variance MolDs or FFs led to good performance, which correspond to a removal variance threshold value of <0.0001 (Supplementary Fig. 8c,d).

Construction of the 2D molecular Fmaps. Upon selection of specific MolD classes or FF sets, their pairwise distances were extracted from the MolMap distance matrices. Based on these pairwise distances, the MolDs or FFs were projected onto a 2D feature space as feature points by using UMAP[19,20]. These feature points embed the broadly learned correlation relationships of the selected MolD classes or FF sets. They were further assigned to the regular grids of a 2D-grid map, MolMap 2D Fmap, by using the Jonker–Volgenant (J–V) algorithm for linear assignment[58]. Here, the J–V algorithm was used for minimizing the cost squared distance matrix $d\_\text{sqeuclidean}_{(x,y)} = \left\| x^{\text{embed}} - y^{\text{grid}} \right\|^2$ so that the MolMap 2D Fmaps maintains the broadly learned correlation relationships of the MolDs or FFs. MolMap (flowchart in Fig. 1) was coded in Python 3+ (Supplementary Fig. 9).

MolMap multichannel Fmaps. Deep learning performance of some multiframe or multiclass data can be enhanced by using multichannel networks, that is, each distinguished data class is learned through a separate channel of a multichannel CNN architecture (multichannel in input layer or in multiple layers). In MolMap, each of the 13 MolD classes is a unique class of molecular properties, and each of the 12 FF sets is of distinguished substructure encoding system. Therefore, in some cases, MolMap-based deep learning of pharmaceutically relevant properties may also be enhanced by learning each MolD class or FF set in separate channel of multichannel networks. To support multichannel deep learning, MolMap outputs the molecular Fmaps in both single-channel and multichannel mode, where each MolD or FF is located at the same grid point in both modes (Fig. 1a).

MolMapNet deep learning architecture. A dual-path CNN architecture[59], MolMapNet, was constructed for simultaneous learning from both MolDs and FFs. MolMapNet consists of three components, the multichannel input Fmaps, dual-path CNN feature learning and the nonlinear task learning (Fig. 3). In this work, MolMapNet MolRs were based on 13 MolD classes and 3 FF sets (MACCSFP, PharmacoErGFP and PubChemFP). Although each selected FF set has been outperformed by the MorganFP (ECFP-like) in various tasks[53,60,61], the MolMap Fmap of each selected FF set presents distinguished and more densely clustered patterns (Supplementary Fig. 10), which enables more enhanced learning of pharmaceutical properties than the other individual FF sets as tested on the three regression and five classification benchmark datasets (Supplementary Fig. 11). Moreover, each selected FF set has performed well in representing molecular databases[62], machine learning tasks[63–65] and virtual screening[66,67]. Collective use of these three FF sets probably lead to better performances.

The first convolution layer of MolMapNet contains a higher number of kernels (48) with larger size (13 × 13/1) for enhanced expressive capability and perception[34]. The max-pooling layer (3 × 3/2) with stride 2 is used after each convolution layer for lower computing cost. To achieve optimal OOTB performances, MolMapNet adopts the naive inception layer derived from GooLeNet[68], which has three parallel small kernels (sizes of 1 × 1, 3 × 3 and 5 × 5) for enhanced local perception. Subsequently, the global max-pooling layer is used for reduced parameters, followed by two or three dense layers for improved nonlinear transformation capability. MolMapNet can be classified into the MolD-only single-path model (MolMapNet-D), the FF-only single-path model (MolMapNet-F) and the joint MolD and FF dual-path model (MolMapNet-B). The maximum number of parameters are no more than 0.83 million.

MolMapNet OOTB hyperparameters and training. These hyperparameters and default settings are summarized in Supplementary Table 4. The activation function rectified linear unit (ReLU) was used for both classification and regression tasks. A small learning rate (0.0001) and batch size (128) were set for all tasks. Further lowering the batch size (for example, 64, 8) can substantially improve the convergence rate and prediction performance for smaller datasets (Supplementary Fig. 3). The batch size 128 was selected upon balanced consideration of the level of improved performances for smaller datasets and the efficiency for training larger datasets. Other regularization options such as dropout and weight decay were not used because the models were easily trained to convergence. In the regression tasks, the loss function was set to mean squared error. In the classification tasks, the weighted cross-entropy loss[32] was used. The early-stopping strategy was used for model training, which has been extensively used in the GCNs and other deep learning models for reduced over-fitting and computing cost[3,4,27,28,32]. Multitask training can be conducted by either joint training[32] or alternate training[28] methods.

For each multitask benchmark dataset, we used the same training method as that of the published deep learning model we aim to compare with. In joint training, all tasks were simultaneously trained and the gradient was built on a global joint output unit. In alternate training, first a base model was trained by 20 global iterations with the optimizer switched from task to task, then every task was trained based on the base model by using the validation set as a monitor for early stopping. All models were developed by TensorFlow 1.14 on GeForce RTX 2080 Ti (12 GB memory in each card) and repeated three times in different split indices or in random seeds. The training details are provided on GitHub at https://github.com/shenwanxiang/bidd-molmap/tree/master/paper.

Benchmark datasets, performance evaluation and metrics. The performances of MolMapNet were extensively tested on 26 common benchmark datasets (Table 1) in comparison with the published performance of the SOTA deep learning models on the same datasets and data-split (training, validation, test) sets. The benchmark datasets and their data-split sets (Supplementary Table 9) were from seven publications[3,4,6,27,28,32,69] and the released codes. These include 3 physicochemical, 3 molecular binding, 14 bioactivity, 3 toxicity and 3 pharmacokinetic datasets, where 10 and 16 datasets are for classification and regression tasks respectively, and 7 classification and 1 regression datasets are multitask. The compared SOTA deep learning models are GCNs (GCNs in MoleculeNet[3], D-MPNN[5], AttentiveFP[4]), FC-DNNs (AE-based FC-DNN[32], Sanofi-Aventis FC-DNN[28]) and CNNs (KekuleScope[27]). In accordance with the evaluation methods of these models, the regression tasks were evaluated by RMSE or R^2, and the classification tasks were evaluated by ROC-AUC or PRC-AUC. The benchmark datasets of some publications are slightly different from one another, partly because of such reasons as duplicates. In this work, we used exactly the same datasets as the published works we directly compare with (for example, D-MPNN[5] and AttentiveFP[4]).

In MolMap, the missing data points (the true labels) in some multitasks such as Tox21 and MUV are processed by a similar approach as that of DeepChem[70]. Specifically, these missing data points are masked as −1 (the classification task) instead of 0 (inactive), and subsequently ignored when computing the loss. In DeepChem, although the missing data points are masked as 0, an additional weight matrix is introduced for all data points including the missing data points. When computing the loss, the loss is multiplied by the weight matrix such that the missing data points are ignored.

Data availability

The full datasets and corresponding annotations are available on GitHub at https://github.com/shenwanxiang/ChemBench/tree/v0 and on Zenodo at https://doi.org/10.5281/zenodo.4054866[71]. Source data are provided with this paper.

Code availability

Codes for the MolMap and MolMapNet package and the parameters are available on GitHub and CodeOcean, together with the data used for testing the package, at https://github.com/shenwanxiang/bidd-molmap and https://codeocean.com/capsule/2307823/tree[72].

Received: 16 March 2020; Accepted: 22 January 2021;
Published online: 01 March 2021

References

1. Paolini, G. V., Shapland, R. H. B., van Hoorn, W. P., Mason, J. S. & Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **24**, 805–815 (2006).
2. Zhavoronkov, A. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
3. Wu, Z. et al. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
4. Xiong, Z. et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* **63**, 8749–8760 (2019).
5. Yang, K. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
6. Duvenaud, D. K. et al. Convolutional networks on graphs for learning molecular fingerprints. *Adv. Neural Inf. Process. Syst.* **28**, 2224–2232 (2015).
7. Moriwaki, H., Tian, Y. S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminform.* **10**, 4 (2018).
8. Landrum, G. RDKit Documentation Release 2019.09.1, 1-151 http://www.rdkit.org (2019).
9. Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
10. Chen, Y., Bai, Y., Zhang, W. & Mei, T. Destruction and construction learning for fine-grained image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition* 5157–5166 (CVPR, 2019).
11. Johnson, R. & Zhang, T. Effective use of word order for text categorization with convolutional neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 103–112 (ACL, 2014).
12. Han, S., Pool, J., Tran, J. & Dally, W. Learning both weights and connections for efficient neural network. In *Proceedings of Advances in Neural Information Processing Systems* 1135–1143 (NIPS, 2015).
13. Dong, J. et al. PyBioMed: a Python library for various molecular representations of chemicals, proteins and DNAs and their interactions. *J. Cheminform.* **10**, 16 (2018).
14. 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).
15. Probst, D. & Reymond, J.-L. A probabilistic molecular fingerprint for big data settings. *J. Cheminform.* **10**, 66 (2018).
16. Allen, C. H. G., Mervin, L. H., Mahmoud, S. Y. & Bender, A. Leveraging heterogeneous data from GHS toxicity annotations, molecular and protein target descriptors and Tox21 assay readouts to predict and rationalise acute toxicity. *J. Cheminform.* **11**, 36 (2019).
17. Clemons, P. A. et al. Quantifying structure and performance diversity for sets of small molecules comprising small-molecule screening collections. *Proc. Natl Acad. Sci. USA* **108**, 6817–6822 (2011).
18. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **4**, 90–98 (2012).
19. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).
20. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. Preprint at https://arxiv.org/abs/1802.03426 (2018).
21. Yao, Q. et al. Taking human out of learning applications: a survey on automated machine learning. Preprint at https://arxiv.org/abs/1810.13306 (2018).
22. Sun, M. et al. Graph convolutional networks for computational drug development and discovery. *Brief. Bioinform.* **21**, 919–935 (2019).
23. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4**, eaap7885 (2018).
24. Goh, G. B., Hodas, N. O., Siegel, C. & Vishnu, A. Smiles2vec: an interpretable general-purpose deep neural network for predicting chemical properties. Preprint at https://arxiv.org/abs/1712.02034 (2017).
25. Karpov, P., Godin, G. & Tetko, I. V. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *J. Cheminform.* **12**, 17 (2020).
26. Goh, G. B., Siegel, C., Vishnu, A. & Hodas, N. O. Chemnet: a transferable and generalizable deep neural network for small-molecule property prediction. Preprint at https://arxiv.org/abs/1712.02734 (2017).
27. Cortés-Ciriano, I. & Bender, A. KekuleScope: prediction of cancer cell line sensitivity and compound potency using convolutional neural networks trained on compound images. *J. Cheminform.* **11**, 41 (2019).
28. 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).
29. Ivan, C. Convolutional neural networks on randomized data. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops* 1–8 (CVPR, 2019).
30. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7**, 20 (2015).
31. Pyka, A., Babuska, M. & Zachariasz, M. A comparison of theoretical methods of calculation of partition coefficients for selected drugs. *Acta Pol. Pharm.* **63**, 159–167 (2006).
32. 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).
33. Ramsundar, B. et al. Massively multitask networks for drug discovery. Preprint at https://arxiv.org/abs/1502.02072 (2015).
34. Peng, C., Zhang, X., Yu, G., Luo, G. & Sun, J. Large Kernel matters—improve semantic segmentation by global convolutional network. In *IEEE Conference on Computer Vision and Pattern Recognition* 4353–4361 (CVPR, 2017).
35. Bento, A. P. et al. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **42**, D1083–D1090 (2014).
36. Probst, D. & Reymond, J.-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminform.* **12**, 12 (2020).
37. Godden, J. W., Stahura, F. L. & Bajorath, J. Anatomy of fingerprint search calculations on structurally diverse sets of active compounds. *J. Chem. Inf. Model.* **45**, 1812–1819 (2005).
38. Fisher, A., Rudin, C. & Dominici, F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* **20**, 1–81 (2019).
39. Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **40**, 773–777 (2000).

40. Tetko, I. V., Tanchuk, V. Y., Kasheva, T. N. & Villa, A. E. P. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* **41**, 1488–1493 (2001).
41. Huuskonen, J., Rantanen, J. & Livingstone, D. Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur. J. Med. Chem.* **35**, 1081–1088 (2000).
42. Huuskonen, J. Estimation of water solubility from atom-type electrotopological state indices. *Environ. Toxicol. Chem.* **20**, 491–497 (2001).
43. Ensing, B. et al. On the origin of the extremely different solubilities of polyethers in water. *Nat. Commun.* **10**, 2893 (2019).
44. Cheng, Y. et al. From fragment screening to in vivo efficacy: optimization of a series of 2-aminoquinolines as potent inhibitors of beta-site amyloid precursor protein cleaving enzyme 1 (BACE1). *J. Med. Chem.* **54**, 5836–5857 (2011).
45. Madden, J. et al. Fragment-based discovery and optimization of BACE1 inhibitors. *Bioorg. Med. Chem. Lett.* **20**, 5329–5333 (2010).
46. Van De Waterbeemd, H. & Gifford, E. ADMET in silico modelling: towards prediction paradise? *Nat. Rev. Drug Discov.* **2**, 192–204 (2003).
47. Altae-Tran, H., Ramsundar, B., Pappu, A. S. & Pande, V. Low data drug discovery with one-shot learning. *ACS Cent. Sci.* **3**, 283–293 (2017).
48. Glavatskikh, M., Leguy, J., Hunault, G., Cauchy, T. & Da Mota, B. Dataset's chemical diversity limits the generalizability of machine learning predictions. *J. Cheminform.* **11**, 69 (2019).
49. Townsend, J., Micucci, C. P., Hymel, J. H., Maroulas, V. & Vogiatzis, K. D. Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nat. Commun.* **11**, 3230 (2020).
50. Kim, S. et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **47**, D1102–D1109 (2019).
51. O'Boyle, N. M. et al. Open Babel: an open chemical toolbox. *J. Cheminform.* **3**, 33 (2011).
52. 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).
53. Capecchi, A., Probst, D. & Reymond, J.-L. One molecular fingerprint to rule them all: drugs biomolecules, and the metabolome. *J. Cheminform.* **12**, 43 (2020).
54. Stiefl, N., Watson, I. A., Baumann, K. & Zaliani, A. ErG: 2D pharmacophore descriptions for scaffold hopping. *J. Chem. Inf. Model.* **46**, 208–220 (2006).
55. McGregor, M. J. & Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Model.* **39**, 569–574 (1999).
56. 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).
57. Sha, Y., Phan, J. H. & Wang, M. D. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society(EMBS)* 6461–6464 (Institute of Electrical and Electronics Engineers, 2015).
58. Jonker, R. & Volgenant, A. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing* **38**, 325–340 (1987).
59. Chen, Y. et al. Dual path networks. *Adv. Neural Inf. Process. Syst.* **30**, 4467–4475 (2017).
60. Skinnider, M. A., Dejong, C. A., Franczak, B. C., McNicholas, P. D. & Magarvey, N. A. Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm. *J. Cheminform.* **9**, 46 (2017).
61. Škuta, C. et al. QSAR-derived affinity fingerprints (part 1): fingerprint construction and modeling performance for similarity searching, bioactivity classification and scaffold hopping. *J. Cheminform.* **12**, 39 (2020).
62. Fernández-de Gortari, E., García-Jacas, C. R., Martinez-Mayorga, K. & Medina-Franco, J. L. Database fingerprint (DFP): an approach to represent molecular databases. *J. Cheminform.* **9**, 9 (2017).
63. Sato, T., Honma, T. & Yokoyama, S. Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening. *J. Chem. Inf. Model.* **50**, 170–185 (2010).
64. Smusz, S., Kurczab, R. & Bojarski, A. J. The influence of the inactives subset generation on the performance of machine learning methods. *J. Cheminform.* **5**, 17 (2013).
65. Czarnecki, W. M., Podlewska, S. & Bojarski, A. J. Robust optimization of SVM hyperparameters in the classification of bioactive compounds. *J. Cheminform.* **7**, 38 (2015).
66. Askjaer, S. & Langgård, M. Combining pharmacophore fingerprints and PLS-discriminant analysis for virtual screening and SAR elucidation. *J. Chem. Inf. Model.* **48**, 476–488 (2008).
67. Venkatraman, V., Pérez-Nueno, V. I., Mavridis, L. & Ritchie, D. W. Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods. *J. Chem. Inf. Model.* **50**, 2079–2093 (2010).
68. Szegedy, C. et al. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition* 1–9 (CVPR, 2015).
69. Mayr, A. et al. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* **9**, 5441–5451 (2018).
70. 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, 2019).
71. Shen, W. X. et al. ChemBench: the molecule benchmarks and MolMapNet datasets. *Zenodo* https://doi.org/10.5281/zenodo.4054866 (2020).
72. Shen, W. X. et al. The molmap package. *Zenodo* https://doi.org/10.5281/zenodo.4056290 (2020).

Acknowledgements

We appreciate the financial support from the National Key R&D Program of China, Synthetic Biology Research (2019YFA0905900), Shenzhen Municipal Government grants (no. 2019156, JCYJ20170413113448742 and no. 201901), Department of Science and Technology of Guangdong Province (no. 2017B030314083) and Singapore Academic Funds R-148-000-273-114. We thank P. J. Feng for his help on providing CYP450 data. We also thank S. Liang and X. Zeng for their help in data analysis and evaluations.

Author contributions

Y.Z.C and W.X.S. designed the study and wrote the manuscript. W.X.S. and X.Z. performed the experiments and data analysis. Y.Y.J. and Y.Z. provided the experimental platform, F.Z., Y.T., C.Q. and Y.W. provided evaluation and suggestions. All authors contributed to the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-021-00301-6.

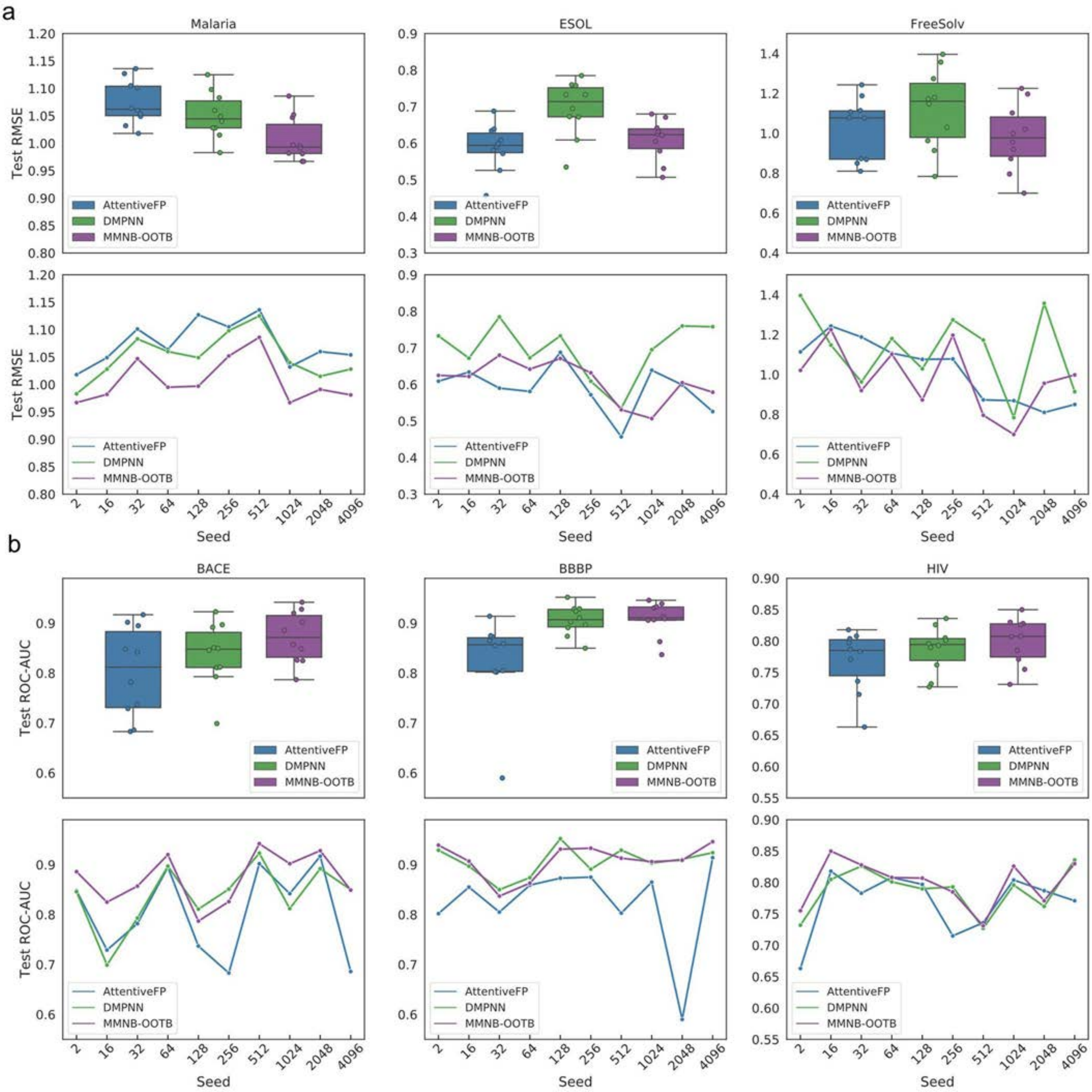
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-021-00301-6.

Correspondence and requests for materials should be addressed to Y.Y.J. or Y.Z.C.

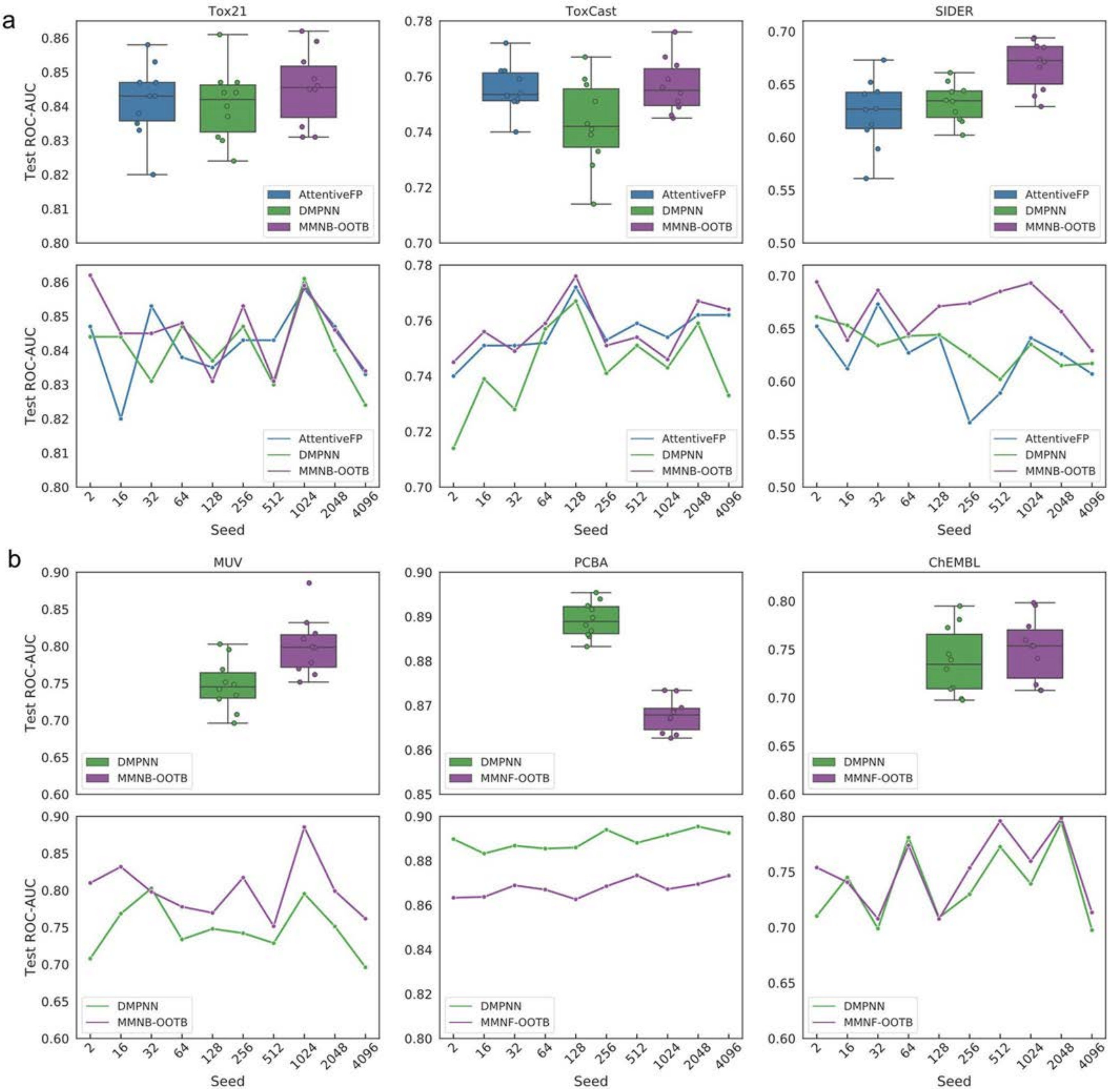
Peer review information *Nature Machine Intelligence* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

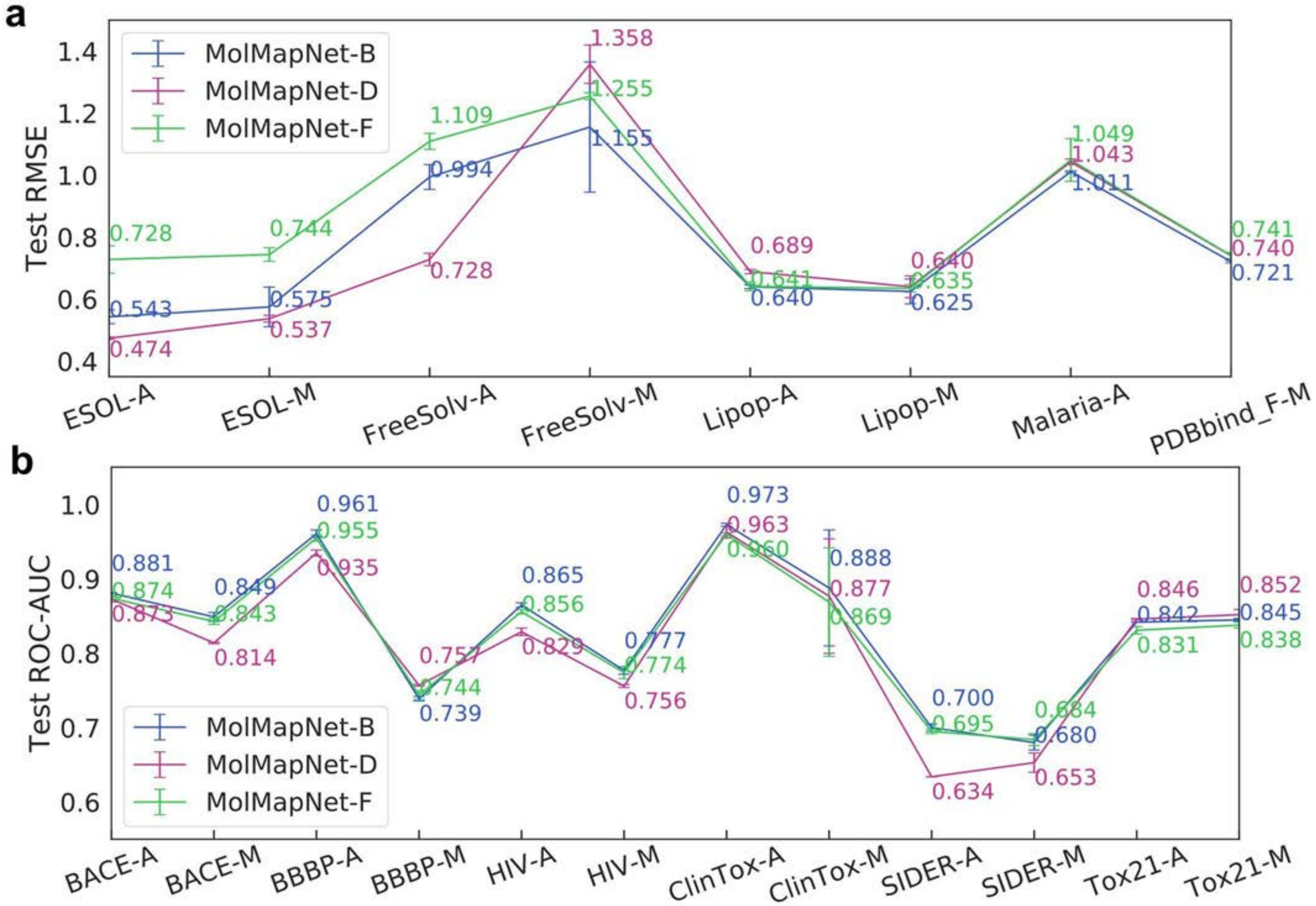
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



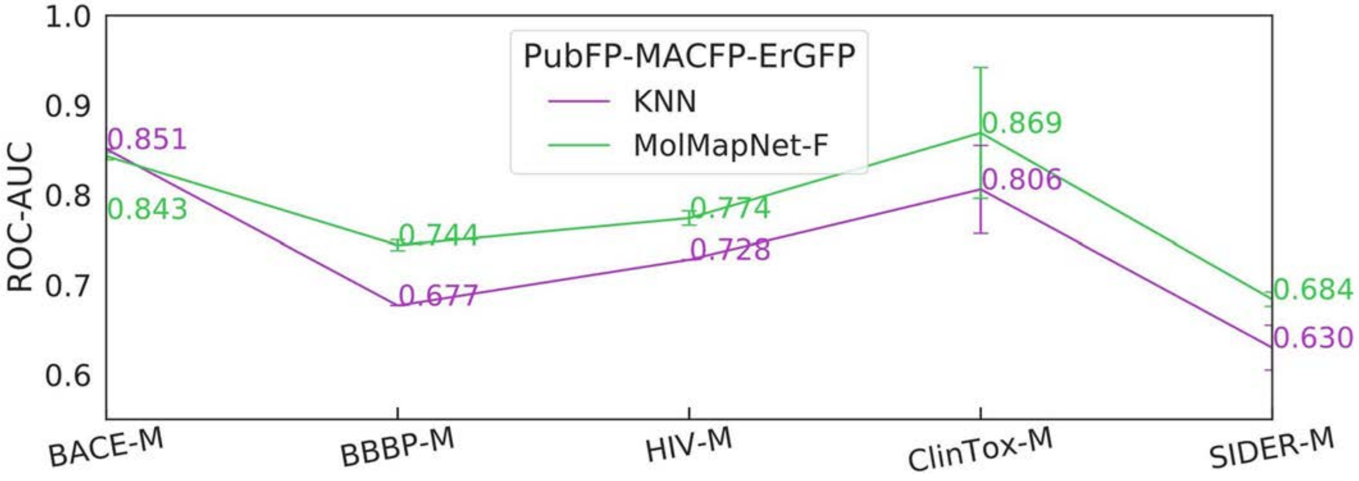
Extended Data Fig. 1 | The performance of the GCN/GAT models and the MolMapNet-OOTB model on 6 single-task benchmark datasets under 10 different splits. MolMapNet-OOTB model is compared to the D-MPNN and AttentiveFP models, the 10 different random seeds 2, 16, 32, 64, 128, 256, 512, 1024, 2048, and 4096 were used for splitting the training set (0.8), validation set (0.1) and test set (0.1). **a**, 3 regression tasks: Malaria, ESOL, FreeSolv under random split. **b**, 3 classification tasks (BACE, BBBP, and HIV) under the scaffold-split.



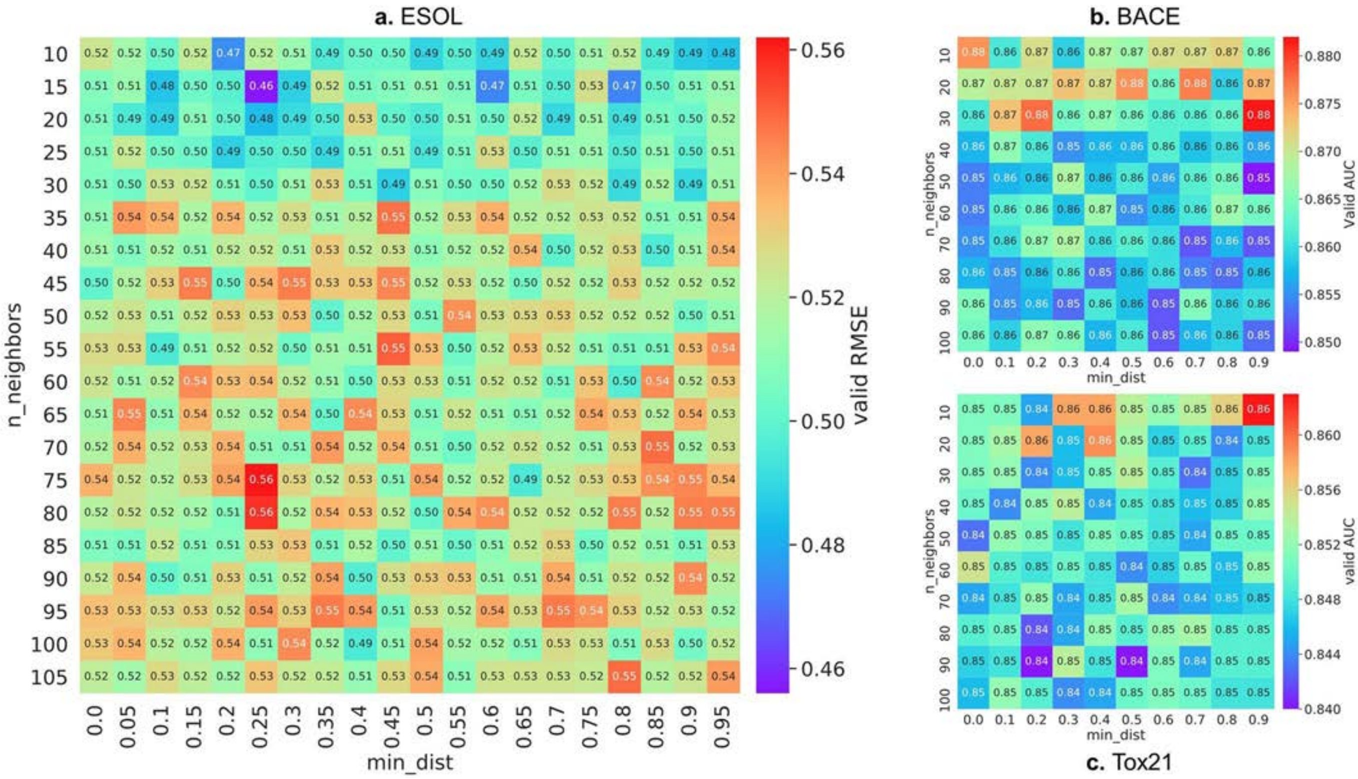
Extended Data Fig. 2 | The performance of the GCN/GAT models and the MolMapNet-OOTB model on 6 multi-task benchmarks under 10 different splits. MolMapNet-OOTB model is compared to the D-MPNN and AttentiveFP models, the 10 different random seeds 2, 16, 32, 64, 128, 256, 512, 1024, 2048, and 4096 were used to split the training set (0.8), validation set (0.1) and test set (0.1). **a**, 3 classification tasks (Tox21, ToxCast, and SIDER) under random split. **b**, 3 high-data classification tasks (MUV, PCBA, and ChEMBL) under random split.



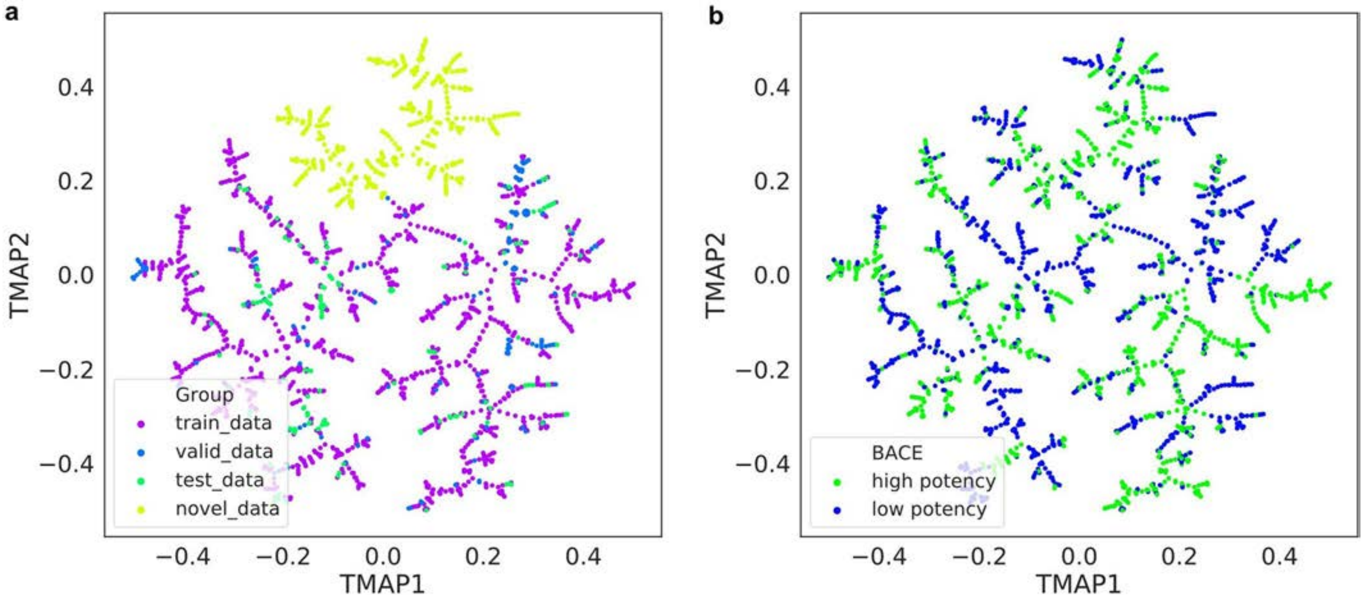
Extended Data Fig. 3 | The performance of single-path MolMapNet-D, MolMapNet-F and dual-path MolMapNet-B models on 11 benchmarks. a, 5 regression benchmark datasets of metric RMSE (ESOL, FreeSolv, Lipop, PDBbind-F, Malaria). **b**, 6 classification benchmark datasets of metric ROC_AUC (BACE, BBBP, HIV, ClinTox, SIDER, Tox21,). These benchmarks are split into training, validation and test set by using both MoleculeNet data-splits (labeled as, for example, ESOL-M) and AttentiveFP data-splits (labeled as, for example, ESOL-A). Note: the error bars represent standard error of the mean.



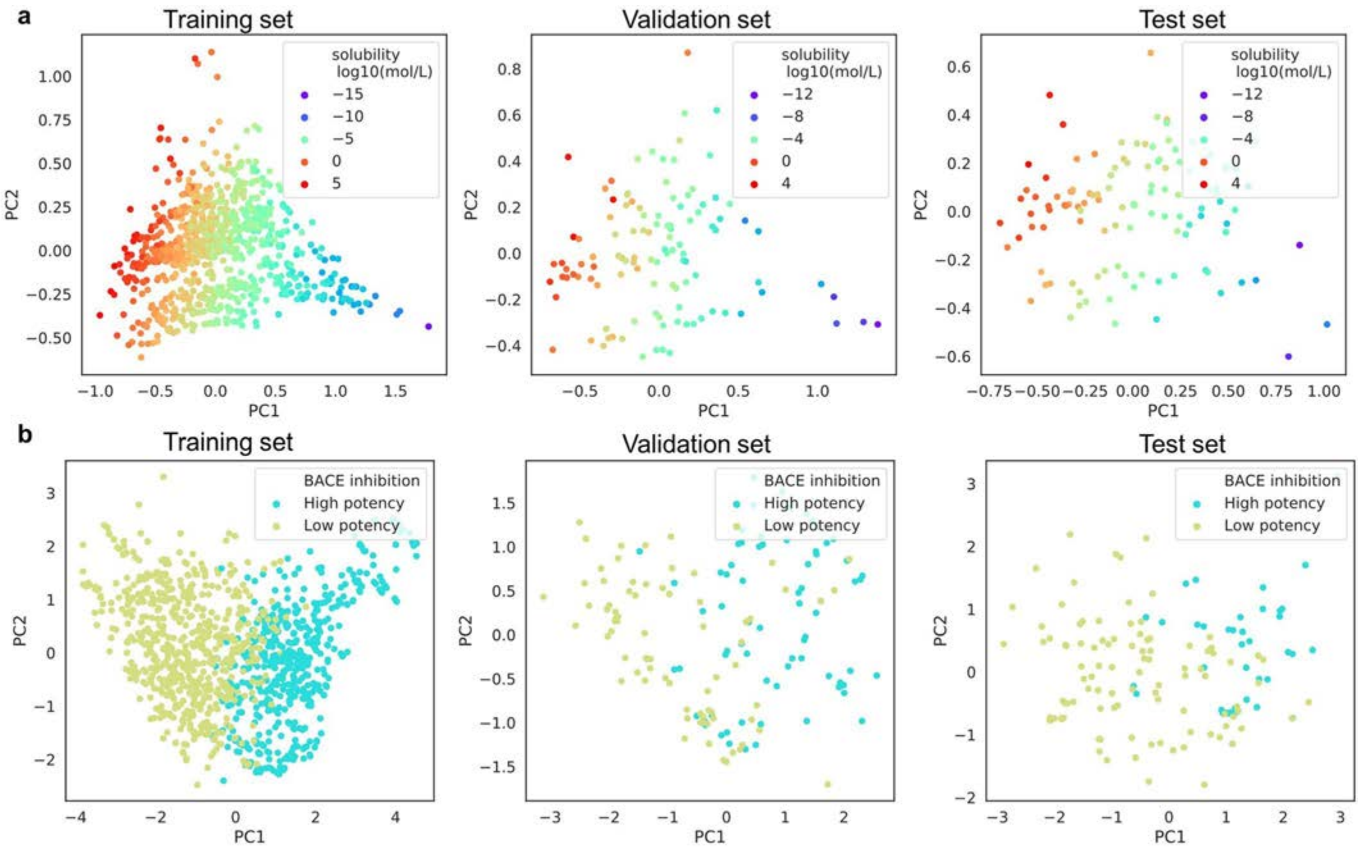
Extended Data Fig. 4 | The performance of kNN and MolMapNet-F on the 5 classification tasks. The 5 classification tasks are under the MoleculeNet data splits, both kNN and MolMapNet-F are based on three sets of fingerprints: PubChemFP, MACCSFP, and PharmacoErGFP (PubFP-MACFP-ErGFP), the error bars represent standard error of the mean.



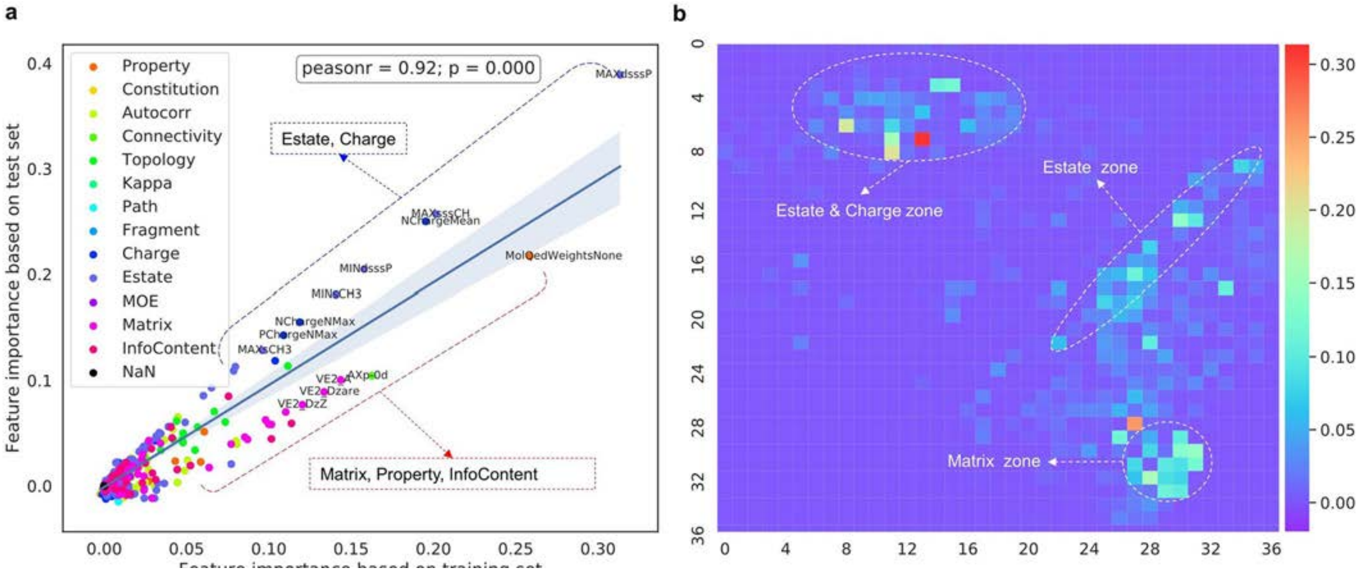
Extended Data Fig. 5 | Optimization of MolMapNet feature-generation parameters n_neighbors and min_dist using grid-search strategy. The parameters n_neighbors and min_dist are in the range of 10~105 and 0~1 respectively, the three datasets ESOL, BACE, and Tox21 are split by the MoleculeNet data-splits method. **a**, optimization of MolMapNet-D model on the ESOL dataset, the performance was evaluated by RMSE of the validation set. **b**, optimization of MolMapNet-F model on the BACE dataset, the performance was evaluated by ROC-AUC of the validation set. **c**, optimization of MolMapNet-B model on the Tox21 dataset, the performance was evaluated by ROC-AUC of the validation set.



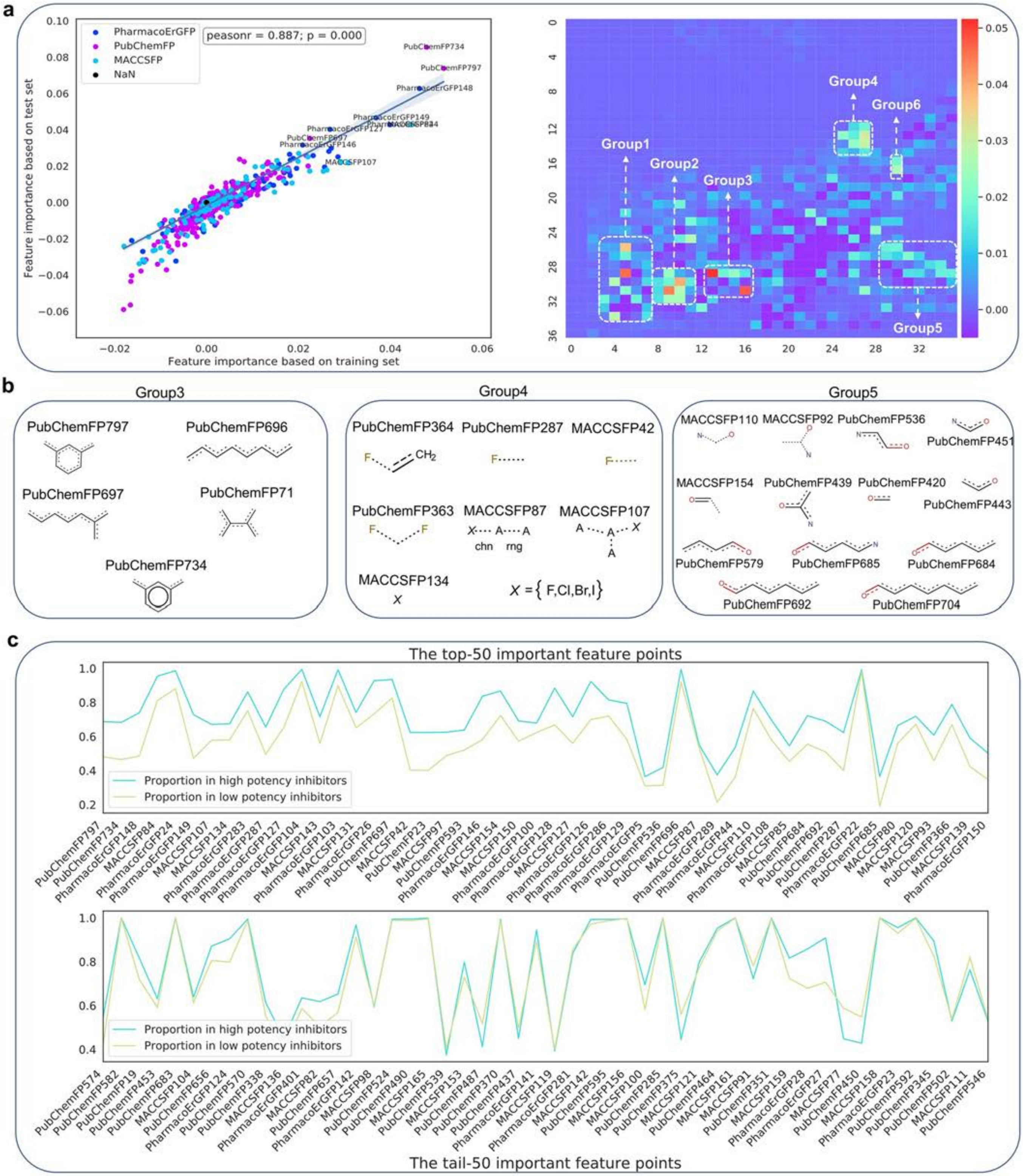
Extended Data Fig. 6 | The TMAP visualization of the BACE training, validation, test and the novel ChEMBL set represented by the 1024-bit Morgan fingerprint(r=2). a, the similarity distribution of the four sets in different color by TMAP[36]: the train_data, valid_data and test_data are the training (646 high potency inhibitors, 564 low potency inhibitors), validation (77 high potency inhibitors, 74 low potency inhibitors), and testing (50 high potency inhibitors, 102 low potency inhibitors) set split from the BACE benchmark dataset using the scaffold-split method, the novel_data is the novel ChEMBL set (216 BACE high potency inhibitors, 179 low potency inhibitors from the ChEMBL database). **b**, the distribution of the compounds with respect to activity type (BACE high potency inhibitors in green and low potency inhibitors in blue color), the interactive visualization is provided at: http://bidd.group/molmap/BACE/BACE.html.



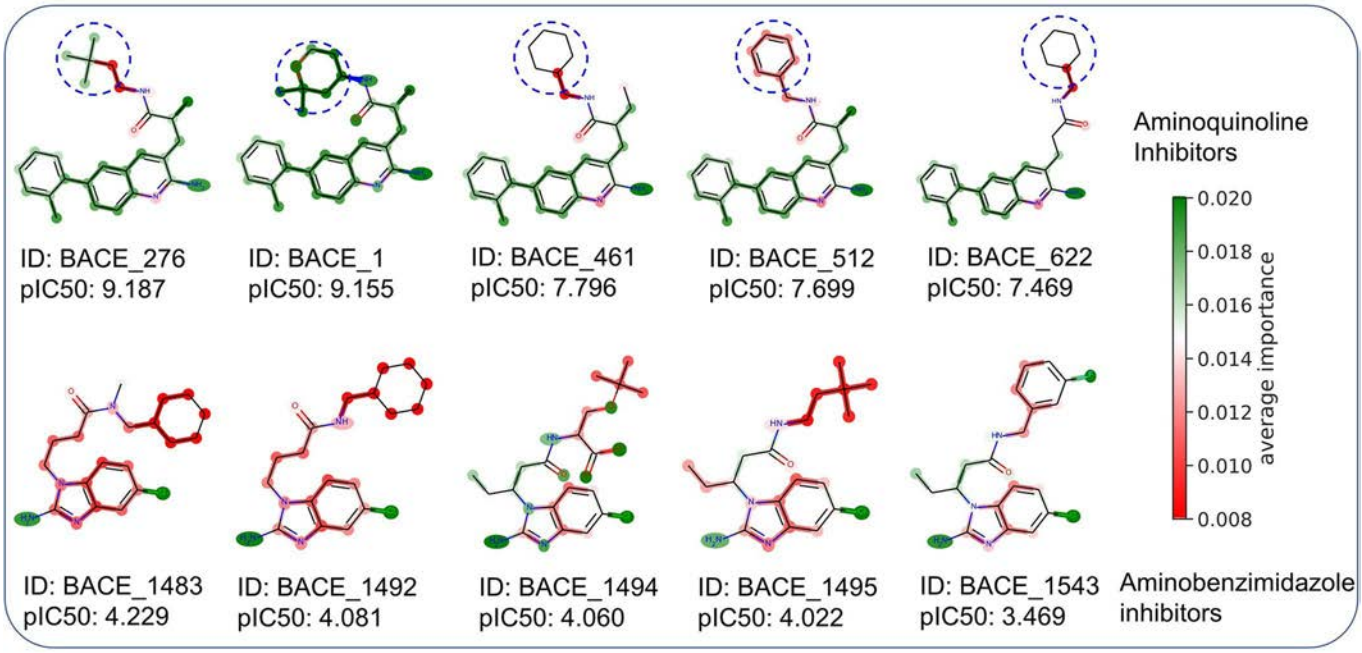
Extended Data Fig. 7 | The PCA of the latent features of the global max pooling (GMP) layer of the MolMapNet-D solubility model and the MolMapNet-F BACE inhibitor model. a, the MolMapNet-D solubility model. **b**, the MolMapNet-F BACE inhibitor model. The MolMapNet-D solubility model was trained on the ESOL benchmark dataset using the AttentiveFP data-split. The MolMapNet-F BACE benchmark model was trained on the BACE dataset using the AttentiveFP data-split (scaffold split).



Extended Data Fig. 8 | The important input-features of the MolMapNet-D solubility model trained on the ESOL dataset using the AttentiveFP data-split. a, the feature importance score of the important features for the ESOL training vs. the test set. **b**, the attention map (the heatmap of the feature importance value). Features of higher positive scores are of higher importance. Features of negative score adversely affect model performance. The top important features are Estate, Charge, Matrix and several other descriptors concentrated in the specific red, orange, and bright green regions in **b**.



Extended Data Fig. 9 | The important input-features of the BACE inhibitor classification MolMapNet-F model. a, the feature importance score of the important features for the BACE training vs. test set (the Pearson correlation coefficient between the two sets is 0.887). and the model attention map (the heatmap of the feature importance value, the smarts patterns of the fingerprint features in the six annotated groups are provided in Supplementary Table 8). **b**, the three groups of the important fingerprints. c, the proportion of the top 50 important features and the bottom 50 features in the BACE high potency and low potency inhibitors.



Extended Data Fig. 10 | The average importance of the atoms and bonds of the BACE inhibitors of two molecular scaffolds in the BACE benchmark dataset. The two molecular scaffolds of BACE inhibitors are 2-aminoquinoline[44] and 2-aminobenzimidazole[45], the atoms and bonds of each inhibitor are color-highlighted based on the presence of 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. The substructures in the dotted circles are consistent with literature-reported structure-activity relationships of BACE inhibitors in previous study[44].