Journal of Cheminformatics

RESEARCH Open Access

Application of 3D atom pair map in an attention model for enhanced drug virtual screening

Gina Ryu[1] and Wankyu Kim[1,2*]

Abstract

Machine learning and artificial intelligence (AI) are actively applied in drug discovery, such as virtual screening, wherein appropriate molecular representation is critical. Conventional compound representations have limited use because they cannot encode the 3D spatial arrangement of atoms. An atom pair map (APM) represents a compound using a numerical matrix that encodes the physicochemical properties of all atom pairs and interatomic distances. In this way, APM inherently captures the 3D shape of a compound, whereas other conventional representations do not, such as fingerprints, SMILES, and molecular graphs. In this study, we performed a step-by-step evaluation of (i) how well APMs encode common molecular characteristics shared among ligands for target or phenotypic screening hits and (ii) how our APM-based attention model (APNet) compares with other conventional and advanced models. We demonstrated that APM and APNet consistently outperformed other representations and related models across various benchmarks.

Scientific contribution This study demonstrates the utility of a novel molecular representation, 3D APM and a deep learning model based on it for virtual screening, suggesting that many other prediction models would also benefit from adopting APM. An open-source script to generate 3D APM is available at https://github.com/rimeless/APM

Keywords Fingerprint, Virtual screening, Structure representation, Deep learning, Drug discovery

Introduction

In silico drug virtual screening enables the efficient identification of hits, sparing the need for high-throughput screening against large compound libraries [1]. Virtual screening can be divided into two types depending on the accessibility of the target structure: ligand-based virtual screening (LBVS) and structure-based virtual screening (SBVS) [2]. LBVS captures similarities between known ligands. It requires different models for each target and becomes less effective for novel targets of few known ligands [3], whereas SBVS is applicable to a broader range of targets by identifying general features responsible for drug-target interactions [4]. Although 3D structures are not always available for certain targets, this limitation can be significantly alleviated by both experimental [5] and in silico modeling [6, 7].

Artificial intelligence (AI) has recently been actively applied to drug discovery. This is largely driven by the application of deep neural networks (DNN) to important problems such as target discovery, in silico drug screening, and the prediction of biochemical properties (*e.g.*, ADME, toxicity) [1]. Virtual screening is one of the most rapidly advancing research areas. The first critical step of this process typically involves encoding the compounds and their respective targets as numerical

*Correspondence:
Wankyu Kim
wkim@ewha.ac.kr
[1] Department of Life Sciences, College of Natural Science, Ewha Womans University, Seoul 03760, Republic of Korea
[2] KaiPharm, Seoul 03759, Republic of Korea

vectors. For example, Rifaioglu et al. embedded 2D images of compounds using a 2D convolutional neural network (2D-CNN) [8], while Ragoza encoded the 3D atom density of compounds using a 3D-CNN [9].

Ideally, a vector representation of compounds should capture two critical pieces of information required for drug-target binding, *i.e.*, the physicochemical properties and the spatial arrangements of atoms. To this end, compounds have been represented as fingerprints [10], molecular graphs [11], atom pair maps (APM) [12]. Fingerprints lack spatial information because they only encode the presence (1 s) or absence (0 s) of a compound's substructures, not their geometric positions. A molecular graph typically treats a compound as a network of atoms, where each atom with its neighbors is encoded as a node vector, and all node vectors are eventually integrated into a single vector. There are several studies related to this, such as those by Lim [13], Zheng [14], and Tsubaki [15]. Molecular graphs still cannot capture the spatial information between distant atoms, although neighboring atoms are encoded during the generation step of the node vectors. In contrast, APMs do not suffer from such limitations because they encode all possible interatomic distances in a compound and the physicochemical properties of the corresponding atom pairs. In an APM, each atom pair is encoded as a combination of physicochemical properties and their interatomic distance (*e.g.*, H-bond donor-hydrophobic at 7 Å, aromatic-acidic at 12 Å, etc.). The interatomic distances can be measured as actual 3D distances [16] or as the minimum number of edges in a molecular graph, as in the early versions of 2D-APM [12]. Here, we developed a more elaborate version of the APM (550 bits) than the earlier versions (16–200 bits) [16]. Notably, we applied APM to protein targets by taking atom pairs around protein-binding sites or pockets. This aspect distinguishes this work from other related studies, which were based on amino acid/peptide compositions [17, 18] and graph-based pocket features [9]. One advantage of our method is that it captures the spatial information of both compounds or targets, whereas the resulting features are rotation-invariant, making the APM highly compatible with any DNN or machine learning model.

Methods

Compound and protein datasets

The 3D atomic coordinates of all the compounds were obtained from PubChem [19]. Compound identifiers were unified, where necessary, by converting SMILES, InChI, or other identifiers using PubChem ID.

Protein 3D structures were obtained from the Protein Data Bank [5] and PDBbind [20]. We use the method of finding binding sites or pockets by Saberi et al. [21]. In brief, a convex hull around the target protein was created using atomic triangles on the surface. The triangles serve as potential pockets and are further filtered and refined by identifying empty voxels and combining overlapping pockets. If the pockets were larger than a $10\times10\times10$ Å^3 cubic unit, they were split into multiple smaller pockets. The pockets were then clustered using K-means, and the centroid of each cluster was selected. Pockets with more than 100 feature pairs were filtered out.

Table 1 Datasets used for training, testing, and validation

	Interaction	Compound	Target	active	decoy
BindingDB	51,579	42,212	714	28,161	23,418
PDBbind	1301	1121	504	806	495
ChEMBL	166,642	101,263	1212	127,592	39,050

Compound-target interaction dataset

The pre-processed dataset of BindingDB [22] by Yingkai et al. [23] was obtained as compound-target interaction dataset. The dataset was divided into training, validation, and test sets. Additionally, the test set was further subdivided into *seen* and *unseen* categories based on presence or absence of the same targets in the training set. We filtered out compounds with highly diverse conformations or molecular weights that were too small or large, resulting in 51,579 compound-target pairs (Table 1).

The bioassay dataset was obtained from ChEMBL [24] by selecting assays of the 'binding' and 'functional' types. We used assays related only to *Homo sapiens*, *Rattus norvegicus*, *Mus musculus*, *Bos taurus*, *Cavia porcellus*, *Sus scrofa*, *Oryctolagus cuniculus*, *Canis familiaris*, *Equus caballus*, *Ovis aries*, *Cricetulus griseus*, *Mesocricetus auratus*, and *Macaca mulatta*.

The assayed compounds were labeled as *active* when pChEMBL > 6.0, where pChEMBL = −log10 (molar IC50, XC50, EC50, AC50, Ki, Kd, or Potency). Compounds with pChEMBL < 4.0 were considered *inactive*. We also filtered out compounds with contradictory labels for both *active* and *inactive* compounds in related assays.

A high-quality subset of PDBbind (ver.2020) was prepared for validation. Compounds with $K_d \leq 10^{-8}$ were labeled as *active* and those with $K_d \geq 10^{-4}$ were labeled *inactive*. This subset was also used to generate APMs for protein pockets.

Generation of the atom pair net (APNet) model

The APNet model consists of four modules, where the APMs of a compound and a target with one or more pockets were taken as input, and their interaction score

was given as an output (Fig. 1). First, the *APM generation module* generates APMs for the compounds and pockets. As a target protein may have one or more pockets, corresponding APMs are generated for each pocket.

Second, the *Embedding module* uses the APMs and generates a combined feature matrix (atom-type pair × distance bins) using CNN1d. We constructed two layers of CNN1d in which a one-dimensional kernel was applied across the distance bins for each pair. The feature matrix was then generated using batch normalization, a rectified linear unit (ReLU), and adaptive max pooling. This procedure was performed for all pockets and compounds.

Third, the attention weight between the compounds and pockets was determined in the *Interaction module*. The type pairs of APM were generated independently of each other, intending to capture potential dependencies among them. To this end, we employed BiLSTM rather than LSTM so that both forward and backward directions were considered. Because each pocket contributes differentially to compound binding, we adopted a multi-head attention algorithm to determine the weights of the different pockets.

Finally, the *Task module* uses the output of the interaction module to calculate the prediction score of the compound-target interaction via a two-layered perceptron, with ReLU as an activation layer and binary cross-entropy loss.

Results and discussion

Generation of atom pair maps (APMs)

APMs were generated as a numerical matrix, where each row and column represent the atom pair type and the binning of the interatomic distances, respectively. Atom pair types were defined by their pairwise combinations of physicochemical properties, such as hydrogen bond *donors* and *acceptors*, *cations*, *anions*, *halogen*, *hydrophobic*, and *aromatics* using SMARTS pattern (Additional file 1: Table S1). If none of the features were assigned using RDkit, we assigned carbon (C) as *hydrophobic*; O, N, and S atoms as *polar*; and the remaining atoms as *others*. If an atom was assigned multiple features simultaneously (*e.g.*, a carbon atom was assigned as both *aromatic* and *hydrophobic*), a single feature was assigned based on the feature priority (donors/acceptors > polar and aromatics > hydrophobic). When an atom acts as both a donor and an acceptor, a composite feature called *donor–acceptor* was indicated, resulting in 10 atom types. Interatomic distances were assigned to 10 exponential bins after inspecting the distance distribution sampled from > 1 million compounds in our dataset (Additional file 1: Figure S1, Methods). Therefore, each APM consisted of a 550-dimensional vector of 55 atom pair types × 10 distance bins.

APM generation consists of three steps (Fig. 2): (i) using a 0's APM matrix, (ii) designating 1's for the corresponding pair type and distance binning for each atom pair in the APM, and (iii) applying Gaussian binning with

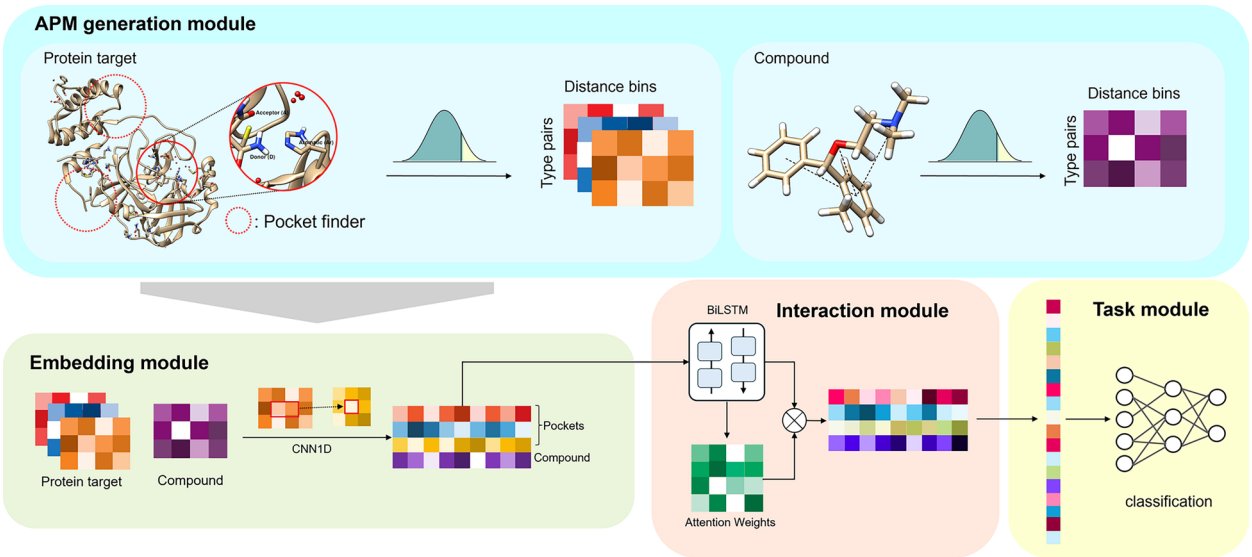
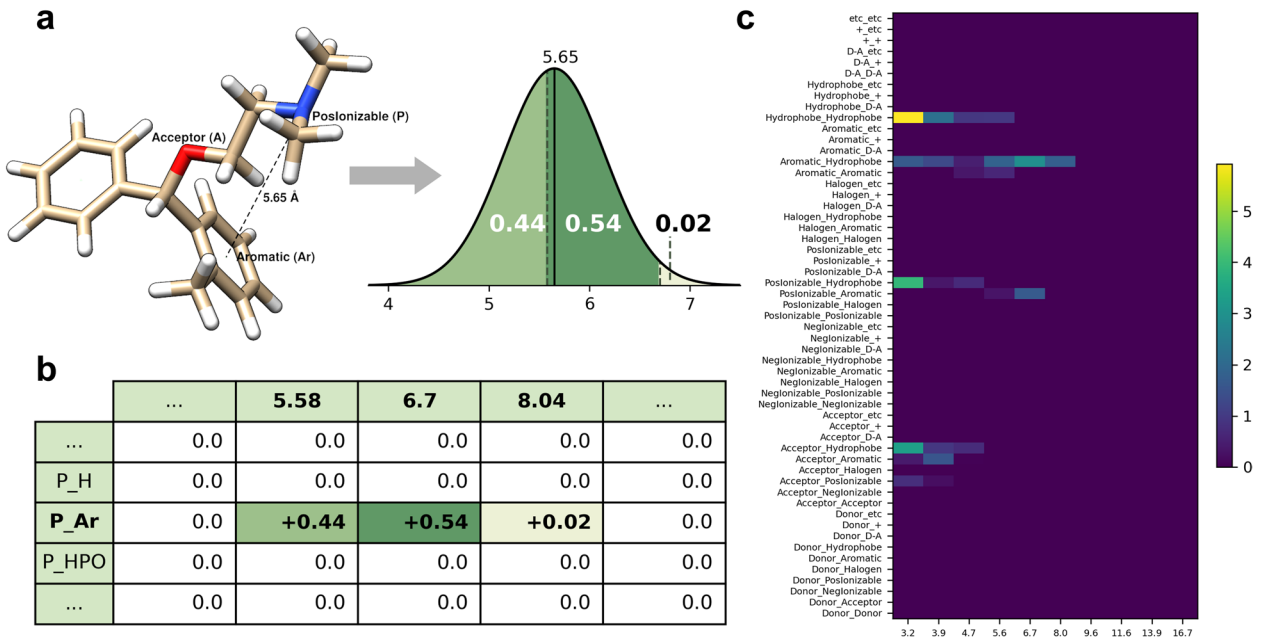


Fig. 1 Generation of the APNet model. (i) Generation of APMs for targets (pockets) and compounds, (ii) embedding APMs for targets and compounds using CNN1d, (iii) bidirectional LSTM learns harmonized characteristics of elements of APMs, and (iv) the self-attention model is trained to determine the pocket weights



	...	5.58	6.7	8.04	...
...	0.0	0.0	0.0	0.0	0.0
P_H	0.0	0.0	0.0	0.0	0.0
P_Ar	0.0	**+0.44**	**+0.54**	**+0.02**	0.0
P_HPO	0.0	0.0	0.0	0.0	0.0
...	0.0	0.0	0.0	0.0	0.0

Fig. 2 Steps in generating atom-pair maps (APMs) **a** Feature extraction from a compound or the target pocket of a protein (*e.g.*, diphenhydramine). **b** Utilizing Gaussian binning on the distance between the extracted feature pairs to distribute values to the corresponding bin and its neighboring bins. **c** Formation of the final APM

a standard deviation of 0.5 to smooth values towards neighboring bins. When multiple atom pairs are mapped to the same type and distance bin, their counts are added cumulatively. Therefore, the sum of APM equals to the total number of atom pairs for the ligand or the pocket.

The APM as an alternative molecular representation

Next, we comparatively evaluated how the APMs can discriminate known ligands from non-ligands relative to other by different molecular representations. Six different methods were compared including APM and other methods such as fingerprints (ECFP4/6 and MHFP6 [25]), molecular graphs (graph2vec [26]) and ErG [27]. As a benchmark, we applied a simple similarity search and ligand-decoy (LD) set collected from BindingDB and ChEMBL. Initially, the total number of targets were 714 and 1,212 in BindingDB and ChEMBL, respectively. In condition I, the targets of less than five *actives* & *inactives* were filtered out, resulting in 179 and 421 targets, respectively. Next, we further removed similar *active-active* pairs of Tc < 0.4 by ECFP4, the remaining 175 and 405 targets were taken as condition II (Fig. 3A). The second filtering was applied only to *active-active* pairs, but not to active-inactive pairs in order to make the evaluation more stringent. For each target, we compared the similarities between the known ligands (L) and LD pairs based on the six different representations (Fig. 3B and C). Among the six methods, APMs best discriminated ligands from decoys (Fig. 3B). Notably, ligands with low similarities were better identified by APMs than other methods (Fig. 3C). The performances were moderately influenced depending on the tools for generating 3D compound structures, where APMs by Open Babel [28] show a moderately lower performance than by PubChem or RDkit (Fig. S2).

Further, we evaluated the APMs under LBVS conditions using antibiotic screening and cytotoxic assay data by Wong et al. [29] as benchmarks. The antibiotic screening dataset consists of 39,152 compounds and their activity values (GR80), including 512 hits. The performance of the APMs was evaluated using equivalent random forest (RF) models and compared with that of ECFP and graph-2vec. The cytotoxic effects of the same compounds were measured in three cell lines (HepG2, HSkMC, and IMR-90 cells), where the hit criteria were set as previously reported [29]. The model was trained using 80% of the dataset; the remaining 20% was reserved for independent evaluation. We performed 20 iterations of random splitting for the training set in an 8:1:1 ratio, where the data were used for training, validation, and testing, respectively. We selected the top 10 out of the 20 models in the validation dataset and further evaluated them against the reserved set. To ensure robustness, the entire process was repeated ten times with varied random seed. The APM consistently showed superior performance across all four evaluations in predicting bioassay hits for antibiotic

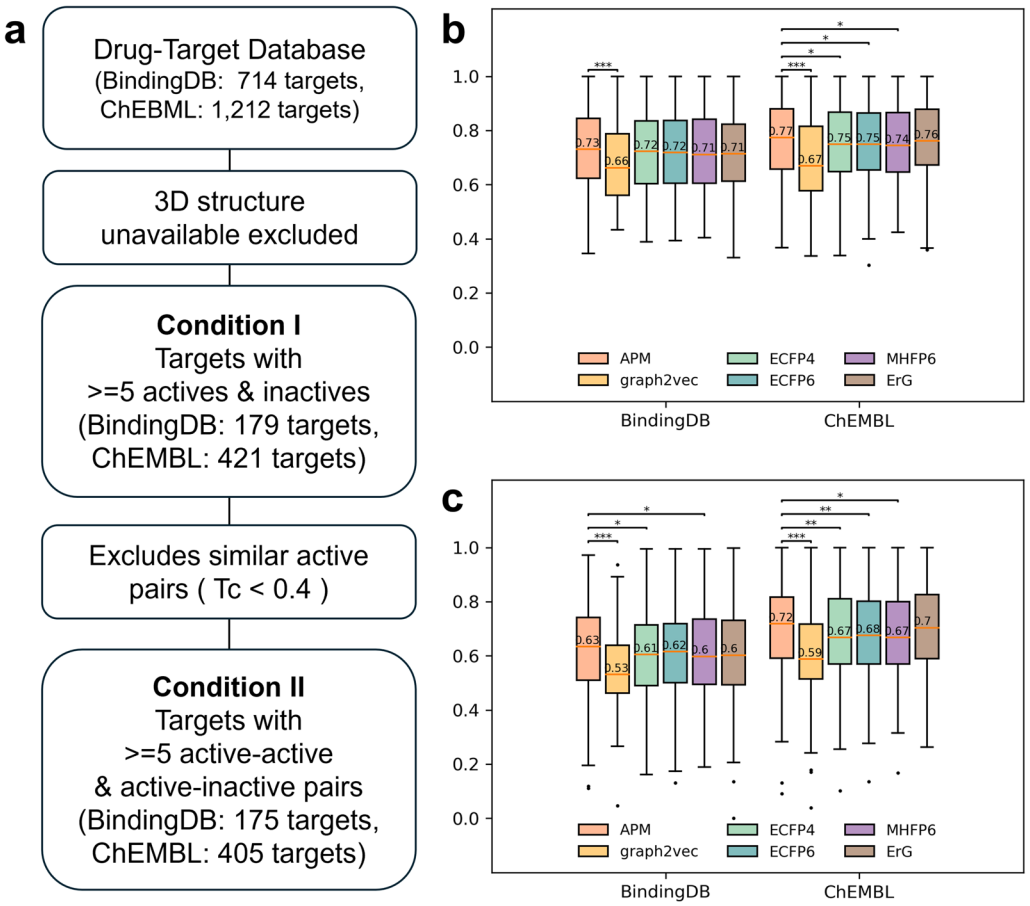


Fig. 3 Performance comparison of the six different molecular representations in discriminating active ligands from inactives using a similarity search. Evaluation scheme of the similarity search using the six molecular representations. **a** The scheme of filtering steps for preparing the benchmark dataset for similarity search. **b** The performance for the targets of 5 or more actives and inactives (Condition I) and **c** the targets after filtering structurally similar active-active pairs (Tanimoto coefficient < 0.4 by ECFP4) (Condition II). The performance was measure as AUROC where active-active pairs were regarded as positives, and active-inactive pairs as negatives. The similarity was measured by Tanimoto coefficient for ECFP4/ECFP6 & MHFP6, weighted Tanimoto coefficient for APM (*i.e.* Ruzicka similarity), cosine similarity for graph2vec and the modified Tanimoto similarity for ErG as in the original study. [27]. (* $p < 0.05$, ** $p < 0.005$ and *** $p < 0.0005$ by t-test)

potency and cytotoxicity compared to the other representations (Fig. 4). In most cases, the APM was significantly better at predicting assay hits than graph2vec or ECFPs. These results indicate that the APM may enhance hit predictions in phenotypic bioassays.

Prediction performance of the APM

To validate the utility of the APM, we compared the performance of different combinations of models and molecular representations. We used the training and test datasets from BindingDB and used PDBbind and ChEMBL for validation. In order to avoid overlap, common data were filtered out among training, test and validation. Additionally, evaluation was conducted in two ways, *seen* and *unseen*, depending on whether the training and the test sets shared the same targets or not, respectively.

First, we constructed four models that used the same RF algorithm but only differed in their molecular representations as input (Table 2). Accordingly, it provides an objective evaluation of their relative performance. The APM showed a slightly better performance in both *seen* and *unseen* tests using BindingDB (Table 3, Fig. 5). Moreover, the APM consistently outperformed the other representations in our validation using the PDBbind and ChEMBL datasets.

Second, we constructed a deep learning (DL) model called APNet using APMs as the input and compared it with four other DL models as benchmarks: DBN [17], drugVQA [14], GNN [15], and AttentionSite [30] (Table 2). Overall, APNet consistently demonstrated the best performance across BindingDB, PDBbind, and

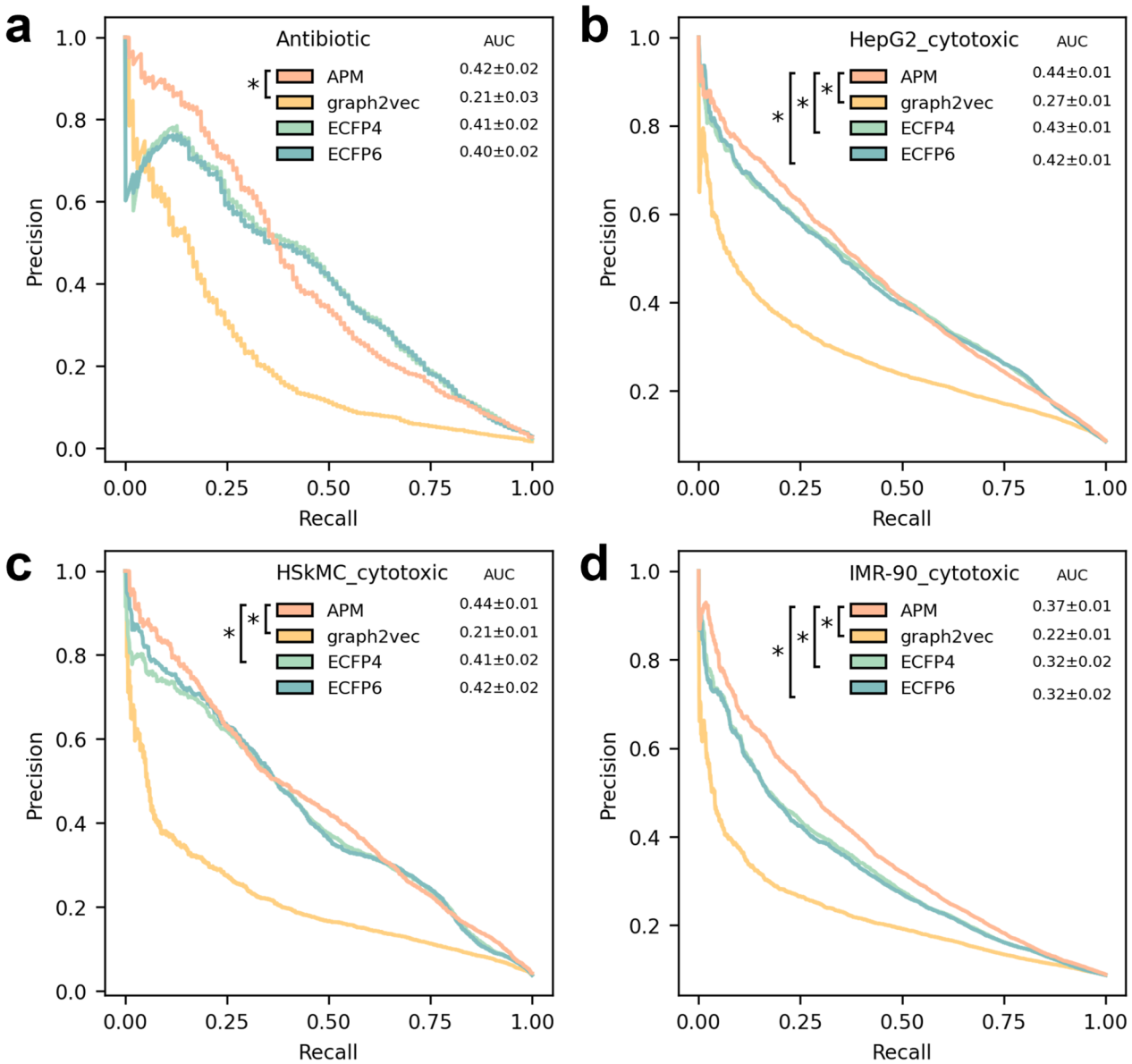


Fig. 4 Calculating the average of the precision-recall curves of APM and the other representations (graph2vec, ECFP4, and ECFP6) using antibiotic efficacy and toxicity bioassay datasets as benchmarks. PR curves were generated by taking the average of 10 iterations of the models × 10 repeats = 100 evaluations. The reserved assay dataset (20%) was used as the benchmark. PR curves using **a** antibiotic assays and cytotoxic assays using the **b** HepG2, **c** HSkMC, and **d** IMR-90 cell lines

ChEMBL, except for the *seen* case of BindingDB (Table 3, Figs. 5–6). APNet showed a relatively better performance with the *unseen* case than the *seen* case in nearly all the cases tested, suggesting that APM-based models may be less affected by overfitting than the other models. APNet was superior to the AutoDock Vina docking model [31] for PDBbind and ChEMBL. Moreover, using PDBbind as a benchmark, the APM-based models (RF_APM and APNet) evidently outperformed all the other models, which is likely to include fewer false positives than BindingDB or ChEMBL. We also checked the influence by the number of bins for RF and APNet. Overall, 10 bins generally show a good performance particularly for the *unseen* case (Figure S3).

Third, we further validated APNet by comparing it with two other DL models, AttentionSite and DBN. Nine screening datasets for the main protease of SARS-CoV-2 were used as benchmarks, which were obtained from PubChem Bioassay (Assay#1,409,579, #1,409,585, #1,409,595, #1,409,599, #1,409,613) [32] and the literature

Table 2 Characteristics of the prediction models

	Ligand type	Target type	Ligand embedding	Target embedding	Model
RF_ECFP	ECFP(0,2,4)	AAC, DAC, TAC	ECFP(0,2,4)	AAC, DAC, TAC	RF
RF_graph2vec	Graph	sequence	graph2vec	Structure2vec (doc2vec base)	RF
RF_APM	APM	APM	APM	APM	RF
DBN	ECFP(0,2,4)	AAC, DAC, TAC	ECFP(0,2,4)	AAC, DAC, TAC	Logistic regression + MLP
GNN	Graph	sequence	GNN(basic)	CNN2D + attention	MLP
drugVQA	smiles	distance map	LSTM + attention	CNN2D + attention	MLP
AttentionSite	Graph	Graph	GCN	GCN	LSTM + attention
APNet	APM	APM	CNN1D	CNN1D	LSTM + attention

Table 3 Performance of the models on the validation databases

	Test		Validation			Summary	
	BindingDB		PDBbind	ChEMBL			
	seen	unseen		seen	unseen	Mean	Std
RF_ECFP4	0.907	0.760	0.621	0.735	0.714	0.747	0.104
RF_gnn	0.971	0.568	0.591	0.551	0.490	0.634	0.192
RF_graph2vec	0.978	0.728	0.689	0.774	0.736	0.781	0.114
RF_APM	**0.988**	0.764	**0.857**	**0.841**	**0.803**	**0.851**	**0.085**
DBN	0.898	0.649	0.534	0.739	0.646	0.693	0.136
GNN	*0.973*	0.862	0.601	0.497	0.510	0.689	0.216
drugVQA	0.956	*0.887*	0.577	0.573	0.514	0.701	0.204
AttentionSite	0.972	0.884	0.650	0.797	0.687	0.798	0.134
APNet	0.946	**0.910**	*0.762*	*0.801*	*0.733*	*0.830*	*0.093*

Boldface text indicates the best-performing model. Italic text indicates t*he second best-performing model*

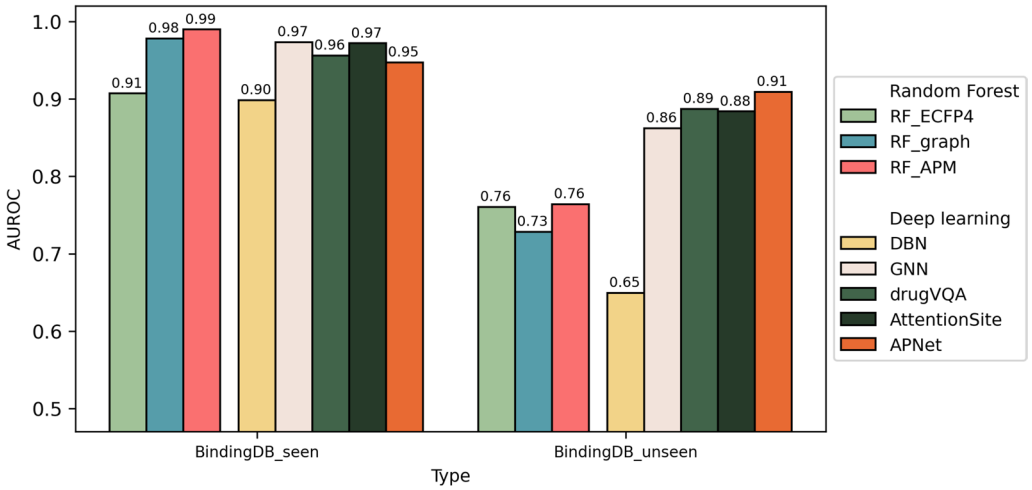


Fig. 5 Comparative performance of APMs and APNet. The AUROC values of these models were benchmarked against other random forest and deep learning models on BindingDB for the seen and unseen protein targets

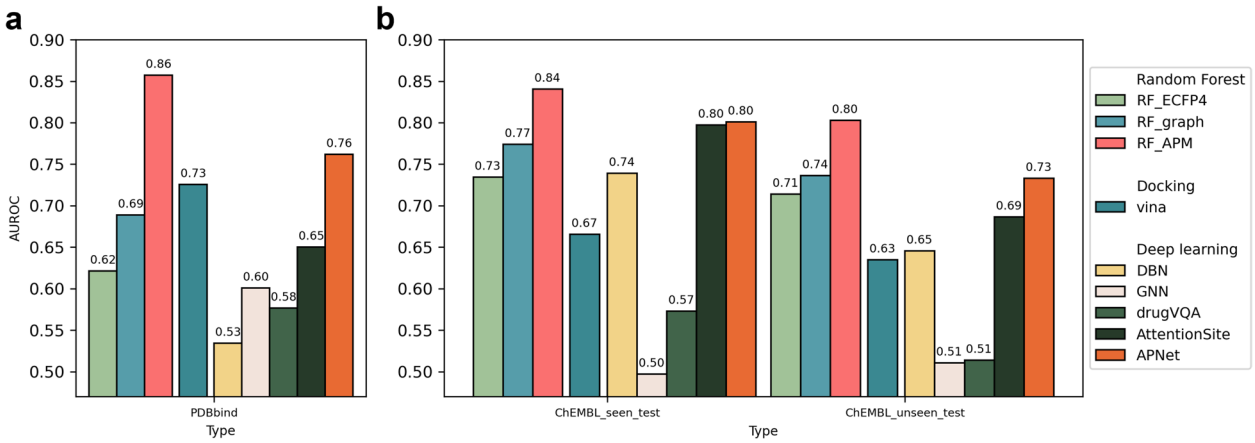


Fig. 6 Validation of the performance of APMs and APNet using external datasets. **a** AUROC scores in PDBbind. **b** AUROC for seen and unseen targets in ChEMBL after training using BindingDB

[33–35]. Ligands are classified to active or decoy against SARS-CoV-2 main protease as reported in assays and literature. All methods employed different but related DL models; however, the main difference lies in the representations of the ligands and targets. APNet is based on APMs, but the other two models used graphs or ECFP/amino acid compositions (Table 2). The results showed that APNet outperformed the two other models in both AUROC and AUPRC in eight of the nine benchmarks (Fig. 7). Considering these results, APM and APNet consistently improved the identification of true ligands or assay hits relative to conventional representations or alternative models, respectively.

Conclusions

In this study, we propose a novel molecular representation, the APM, which has not been widely adopted in rapidly advancing DL-based models. Molecular representations such as SMILES, ECFPs, and graphs have been frequently used for DL-based models, but they only encode 1D or 2D information. APMs have the advantage of encoding physicochemical properties

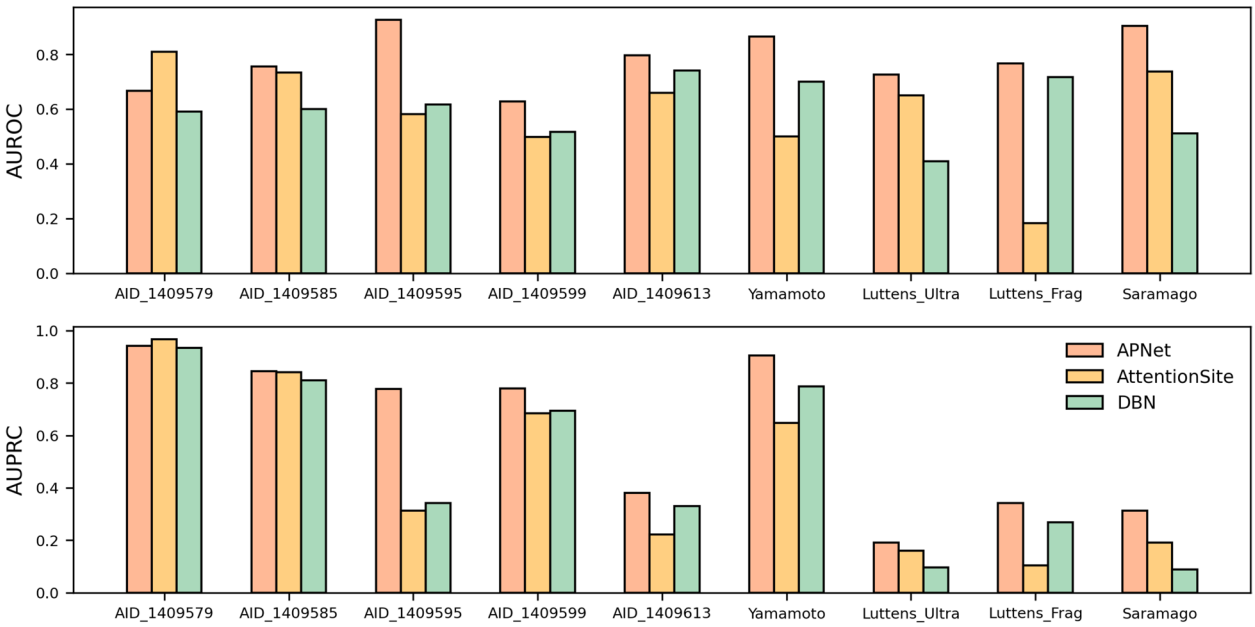


Fig. 7 Model validation via COVID-19 main protease inhibitor screening. Assessment of the predictive performance of the models on PubChem Bioassay Collections

and spatial atomic arrangements in 3D. An ideal molecular representation should capture key features shared among various elements (*e.g.*, a set of ligands) and be generally sufficient to allow discoveries, such as novel scaffolds. We hypothesized that APMs have the potential to improve current molecular representations and performed comparative evaluations using various benchmarks. Using a simple similarity search, we showed that the APM was superior to other representations (*e.g.*, graph2vec and ECFP) in discriminating true ligands. We then constructed APNet, a DL model based on APMs, and demonstrated that it consistently outperformed other related methods across the benchmarks we used. The results suggest that many other prediction models may benefit from adopting the APM as their molecular representation, even without algorithmic modifications. Moreover, the APM is also computationally cost-effective because it has a numerical vector of size 550, which is manageable even with a mid-sized computing facility.

Abbreviations

AI	Artificial intelligence
APM	Atom pair map
AUROC	Area under the receiver-operating characteristic curve
DBN	Deep belief network
DL	Deep learning
DNN	Deep neural networks
LBVS	Ligand-based virtual screening
LD	Ligand-decoy
PR	Precision-Recall
ReLU	Rectified linear unit
RF	Random forest
SBVS	Structure-based virtual screening

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-01023-2.

Additional file1 (PDF 531 KB)

Author contributions

G.N. and W.K. designed and conducted research and wrote the manuscript. Both authors read and approved the final manuscript.

Funding

This work was supported by grants from the National Research Foundation of Korea (NRF), funded by the Korean government (NRF-2022R1A2C1091260 and NRF-2021R1A6A1A10039823).

Data availability

No datasets were generated or analysed during the current study.

Code availability

All codes used in the data analysis and preparation of the manuscript, along with a description of the necessary steps for reproducing the results, can be found in the GitHub repository accompanying this manuscript: https://github.com/rimeless/APM.

Declarations

Ethics approval and consent to participate

The authors declare that they have no competing interests.

Competing Interests

The authors declare no competing interests.

Received: 4 April 2024 Accepted: 24 April 2025
Published online: 05 May 2025

References

1. Kim J, Park S, Min D, Kim W (2021) Comprehensive survey of recent drug discovery using deep learning. Int J Mol Sci. https://doi.org/10.3390/ijms22189983
2. D'Souza S, Prema KV, Balaji S (2020) Machine learning models for drug–target interactions: current knowledge and future directions. Drug Discov Today 25:748–756
3. Riniker S, Landrum GA (2013) Open-source platform to benchmark fingerprints for ligand-based virtual screening. J Cheminform. https://doi.org/10.1186/1758-2946-5-26
4. Lionta E, Spyrou G, Vassilatis D, Cournia Z (2014) Structure-based virtual screening for drug discovery: principles, applications and recent advances. Curr Top Med Chem 14:1923–1938
5. Burley SK et al (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res. https://doi.org/10.1093/nar/gkaa1038
6. Jumper J et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583
7. Baek M et al (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science 373:871–876
8. Rifaioglu AS et al (2020) DEEPScreen: high performance drug-target interaction prediction with convolutional neural networks using 2-D structural compound representations. Chem Sci 11:2531–2557
9. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR (2017) Protein-ligand scoring with convolutional neural networks. J Chem Inf Model 57:942–957
10. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model. https://doi.org/10.1021/ci100050t
11. Clark AM, Labute P, Santavy M (2006) 2D structure depiction. J Chem Inf Model 46:1107–1123
12. Carhart RE, Smith DH, Venkataraghavan R (1984) Atom pairs as molecular features in structure-activity studies: definition and applications. J Chem Inf Comput Sci 25:64–73
13. Lim J et al (2019) Predicting drug-target interaction using a novel graph neural network with 3D structure-embedded graph representation. J Chem Inf Model 59:3981–3988
14. Zheng S, Li Y, Chen S, Xu J, Yang Y (2020) Predicting drug-protein interaction using quasi-visual question answering system. Nat Mach Intell 2:551
15. Tsubaki M, Tomii K, Sese J (2019) Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics 35:309–318
16. Awale M, Jin X, Reymond JL (2015) Stereoselective virtual screening of the ZINC database using atom pair 3D-fingerprints. J Cheminform. https://doi.org/10.1186/s13321-014-0051-5
17. Wen M et al (2017) Deep-learning-based drug-target interaction prediction. J Proteome Res 16:1401–1409
18. Lee I, Keum J, Nam H (2019) DeepConv-DTI: architecture. PLoS Comput Biol 15:e1007129
19. Bolton EE et al (2011) PubChem3D: a new resource for scientists. J Cheminform 3:1–15
20. Liu Z et al (2017) Forging the basis for developing protein-ligand interaction scoring functions. Acc Chem Res 50:302–309
21. Saberi Fathi SM, Tuszynski JA (2014) A simple method for finding a protein's ligand-binding pockets. BMC Struct Biol 14:1–9
22. Gilson MK et al (2016) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res 44:D1045–D1053
23. Yingkai Gao K et al. (2018) Interpretable drug target prediction using deep neural representation. IJCAI International Joint Conference on Artificial Intelligence 2018-July, 3371–3377

24. Davies M et al (2015) ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic Acids Res. https://doi.org/10.1093/nar/gkv352
25. Probst D, Reymond J-L (2018) A probabilistic molecular fingerprint for big data settings. J Cheminform 10:66
26. Narayanan A et al. (2017) graph2vec: Learning Distributed Representations of Graphs. ArXiv
27. Stiefl N, Watson IA, Baumann K, Zaliani A (2006) ErG: 2D pharmacophore descriptions for scaffold hopping. J Chem Inf Model 46:208–220
28. O'Boyle NM et al (2011) Open babel: an open chemical toolbox. J Cheminform 3:33
29. Wong F et al (2024) Discovery of a structural class of antibiotics with explainable deep learning. Nature. https://doi.org/10.1038/s41586-023-06887-8
30. Yazdani-Jahromi M et al (2022) AttentionSiteDTI: An interpretable graph-based model for drug-Target interaction prediction using NLP sentence-level relation classification. Brief Bioinform 23:1–14
31. Trott O, Olson AJ (2009) AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31:455–461
32. Kim S et al (2023) PubChem 2023 update. Nucleic Acids Res 51:D1373–D1380
33. Yamamoto KZ, Yasuo N, Sekijima M (2022) Screening for inhibitors of main protease in SARS-CoV-2. in silico and in vitro approach avoiding peptidyl secondary amides. J Chem Inf Model 62:350–358
34. Saramago LC et al (2023) AI-driven discovery of SARS-CoV-2 main protease fragment-like inhibitors with antiviral activity in vitro. J Chem Inf Model 63:2866–2880
35. Luttens A et al (2022) Ultralarge virtual screening identifies SARS-CoV-2 main protease inhibitors with broad-spectrum activity against coronaviruses. J Am Chem Soc 144:2905–2920

Publisher's Note