

Cite this: *Digital Discovery*, 2025, 4, 3304

# AgreementPred: a cheminformatic framework for drug and natural product category recommendation based on multi-representation structural similarity data fusion

Chayanis Sutcharitchan,  Boyang Wang, Dingfan Zhang, Qingyuan Liu, Tingyu Zhang, Peng Zhang and Shao Li\*

Natural products offer a vast reservoir of bioactive compounds, playing a crucial role in drug discovery. In this big data era, the annotation of their pharmacological categories holds great potential for accelerating drug discovery and advancing mechanistic studies of herbal medicines. However, a vast majority of natural products' classification remains unannotated. Existing recommendation frameworks for pharmacological categories are predominantly tailored to conventional drugs and frequently require extensive experimental data which are typically lacking for natural products. Traditional cheminformatic approaches based on structural similarity, while widely adopted, often struggle to achieve a satisfactory balance between prediction recall and precision, thereby limiting their overall effectiveness. In this study, a simple and explainable category recommendation framework for drugs and natural products based on multi-representation structural similarity data fusion, AgreementPred, was proposed. The framework utilized PubChem compound annotations which comprised two compound classification systems, Anatomical Therapeutic Chemical (ATC) classification and Medical Subject Headings (MeSH) as category labels, extending the scope of application beyond conventional drugs. The similarity search results using 22 molecular representations were combined to improve prediction recall. The predicted annotations were subsequently filtered by agreement scores to enhance prediction precision. Compared to existing equivalent approaches, AgreementPred achieved superior recall-precision balance in both ATC and category prediction tasks. With an agreement score threshold of 0.1, AgreementPred showed 0.74 and 0.55 of recall and precision, respectively, for the category prediction for 1000 compounds from a pool of 1520 categories. Finally, AgreementPred was applied to 321 605 unannotated drugs and natural products. The resulting prediction is expected to be of contribution to drug discovery, as well as mechanistic study purposes.

Received 27th July 2025  
Accepted 29th September 2025

DOI: 10.1039/d5dd00329f

rsc.li/digitaldiscovery

## Introduction

Herbal medicine has been acknowledged as a valuable source for drug discovery, contributing significantly to pharmacological advancement.<sup>1</sup> Natural products isolated from herbal materials have demonstrated clinical efficacy in the treatment of various diseases, with notable examples including ephedrine,<sup>2</sup> artemisinin,<sup>3</sup> and paclitaxel.<sup>4</sup> In recent years, the focus has gradually shifted from screening isolated natural compounds for specific biological activities or targets to exploring the therapeutic benefits of multi-component herbal extracts and formulations, which may offer synergistic pharmacological effects.<sup>5–7</sup> Our team had previously established

a R&D platform, UNIQ system (using Network target for Intelligent and Quantitative analysis on drug actions), focusing on drug discovery and exploration of the holistic mechanism of herbal medicines.<sup>5</sup>

For mechanistic studies, the understanding of the chemical composition of each herb, as well as the pharmacological effects of the components is essential. However, relevant data on natural products remain limited.<sup>8</sup> Unlike synthetic drugs, which benefit from standardized classification systems, a vast majority of natural products' pharmacological classification remains unannotated. Although several databases, such as ChEMBL and the Natural Product Activity and Species Source (NPASS), provide quantitative biological activity data of natural products on specific targets, inferring the classification of a compound solely from biological targets presents a great challenge, particularly given the inherent incompleteness of available datasets.

Institute for TCM-X, Department of Automation, Tsinghua University, 100084 Beijing, China. E-mail: shaoli@mail.tsinghua.edu.cn; Fax: +86 10 62786911; Tel: +86 10 62797035



The Anatomical Therapeutic Chemical (ATC) classification system, established by the World Health Organization (WHO), provides a hierarchical framework for categorizing medical substances based on their anatomical, pharmacological, and chemical properties.<sup>9</sup> As a well-curated and high-quality annotated dataset, it has significantly contributed to the advancement of computational methodologies for predicting new therapeutic applications of existing drugs, thereby facilitating drug repositioning.<sup>10–12</sup>

Inspired by ATC-predicting methods, this study aimed to develop a category recommendation framework that can be applied to both drugs and natural products, using PubChem compound annotations as category labels. PubChem compound annotations comprised two compound classification systems, Anatomical Therapeutic Chemical (ATC) classification and Medical Subject Headings (MeSH). MeSH database, established by the United States National Library of Medicine, provides controlled vocabulary for indexing, cataloging, and searching of biomedical and health-related information.<sup>13</sup> The database curated chemical compounds, including drugs and natural products, related to each MeSH term. Utilizing PubChem compound annotations enabled predictive frameworks to extend its application beyond conventional drug space and provide reasonable annotation for natural products in the database.

Within the domain of natural products, molecular structure remains the most consistently available and reliable source of information for method development. Unlike approved drugs, most natural products lack well-documented data such as chemical–chemical interaction, gene expression, drug target, or side effect profiles utilized in various ATC-predicting methods.<sup>11,12,14–16</sup> Moreover, MeSH terms also lack the hierarchical relationships inherent in the ATC classification system, which several ATC-prediction frameworks have leveraged.<sup>16–18</sup> Therefore, this study focused exclusively on predicting categories using only molecular structures.

In computational chemistry, a molecular structure can be represented in multiple ways, each capturing different aspects of a molecule.<sup>19</sup> Molecular fingerprints are typically employed to represent predefined structural features, such as topological distance between atom pairs, atomic environment within a preset radius, or the presence of specific pharmacophores. Notable examples include atom pair fingerprint (AP), extended connectivity fingerprint (ECFP), and pharmacophore fingerprint (PHFP).<sup>19,20</sup> On the other hand, for deep learning implementation, learned representation based on graph neural network has increasingly gained prominence owing to its flexibility, task-specificity, and oftentimes superior prediction performance compared to predefined molecular descriptors, especially on large datasets.<sup>21–23</sup>

However, to date, there has not been a single molecular representation that outperformed others in all types of tasks and datasets. Previously published ATC prediction frameworks that relied solely on molecular structure as input employed distinct molecular representations.<sup>11,24,25</sup> Yang *et al.*<sup>22</sup> discovered that given certain conditions such as small (less than 1000 molecules) and highly imbalanced dataset, models that

integrated learned representation with fixed molecular descriptor outperformed those that employed only learned representation. Furthermore, Boldini *et al.*<sup>20</sup> investigated the effectiveness of various molecular fingerprints for characterizing the chemical space of natural products, as well as their applicability on the bioactivity prediction. The study revealed inherent variation of pairwise similarity and prediction performance among different molecular fingerprints, highlighting that each fingerprint offered a different aspect of the same molecule.

In this study, the performance of multiple molecular representations, including 28 molecular fingerprints and 1 unsupervised learned representation, in similarity-based category recommendation was further explored on drug and natural product datasets. Moreover, leveraging the integration of multi-representation structural similarity data, a novel category recommendation framework, AgreementPred, was proposed. After eliminating redundant representations, the framework combined the similarity search results of 22 molecular representations and subsequently filtered the predictions using agreement scores. AgreementPred achieved recall-precision balance superior to previous ATC-predicting frameworks in both ATC and category recommendation tasks and was applied to 321 605 unannotated compounds from drug and natural product databases. A total of 2 888 927 categories were recommended for 321 596 compounds with agreement score higher than 0.1. The resulting prediction is expected to be useful in furthering drug discovery, as well as mechanistic study of herbal medicine and natural products.

## Material and methods

### Data collection and preparation

**Datasets.** The aim of this study is to utilize existing classification annotations of drugs and natural products to reasonably predict categories for unannotated natural products. Therefore, compounds of interest in this study comprised those from established databases of modern drugs and natural products, namely DrugBank,<sup>26</sup> SIDER,<sup>27</sup> LOTUS,<sup>28</sup> NPASS,<sup>29</sup> HERB2.0,<sup>30</sup> and TM-MC2.0<sup>31</sup> with collectable PubChem Compound ID (CID). The scope of each database and data from each database used in this study is explained in Table S1.

PubChem record of each compound was obtained by searching concatenated CID lists on PubChem database. The resulting tabular data were composed of names, synonyms, identifiers, chemical properties, and annotations of the compounds. A total of 331 326 PubChem records were collected, in which 9721 compounds contained classification annotations. The annotated records were extracted to construct Annotated-Compound dataset (Table S2).

The drug side effect (SE) dataset was constructed in a similar manner, by mapping compounds in SIDER database to PubChem compounds. Finally, 1376 compounds with obtainable CIDs were incorporated into the Annotated-SE dataset (Table S3).

To reduce computation burden during method development and validation, a sample dataset, AnnoCom1000 was



constructed by random sampling 1000 compounds from Annotated-Compound. Moreover, DrugBank1000 and NP1000 datasets were also constructed by random sampling from annotated compounds contained in DrugBank and natural products databases, respectively. The purpose of constructing these two datasets was to compare the prediction performance of each representation on drug and natural product space.

**Category labels.** PubChem annotations of each compound contained available ATC and/or MeSH codes and terms. These terms are used as category labels in this study. Most of the annotations contained several ontologies of category in broad to specific order. Each level of category was separated by a character ">", and each system of classification was separated by a character "|". For example, the annotation of rosuvastatin was "D004791 - Enzyme Inhibitors > D019161 - Hydroxymethylglutaryl-CoA Reductase Inhibitors|C78276 - Agent Affecting Digestive System or Metabolism > C29703 - Antilipidemic Agent|D057847 - Lipid Regulating Agents > D000960 - Hypolipidemic Agents > D000924 - Anti-cholesteremic Agents|C471 - Enzyme Inhibitor > C1655 - HMG-CoA Reductase Inhibitor|D009676 - Noxae > D000963 - Antimetabolites|C - Cardiovascular system > C10 - Lipid modifying agents > C10A - Lipid modifying agents, plain > C10AA - Hmg coa reductase inhibitors".

For each compound, the terms contained in the annotations were extracted, stripped of codes, and converted to lower-cased letters. Singular and plural versions of the same terms in the dataset were merged (singular versions were kept, if present), and duplicated terms were eliminated from each record. Finally, the resulting Annotated-Compound dataset contained 54 675 compound-annotation pairs with 1520 unique annotations (Table S4). The sample datasets, AnnoCom1000, DrugBank1000, and NP1000, contained 5612, 6978, 3995 compound-annotation pairs, comprising 872, 971, and 544 unique annotations, respectively.

In this study, minimization of manual manipulation of category labels was intended and rationalized that all unique labels, albeit highly similar, had different positions in the chain of ontology (Table S5) and manual aggregation of the labels could compromise the traceability of the related annotations. For instance, antiparkinsonian agent and antiparkinson agents belong to separate chains of ontology, namely, C78276 - Agent Affecting Nervous System > C38149 - Antiparkinsonian Agent and D002491 - Central Nervous System Agents > D018726 - Anti-Dyskinesia Agents > D000978 - Antiparkinson Agents, respectively. Merging the two terms would obscure distinction between the two ontology systems. In contrast, preserving them as separate terms allows potential connection to be drawn while acknowledging that difference may exist. Thus, several similar labels, such as antidepressant agent and antidepressants and antiparkinsonian agent and antiparkinson agents, were kept as is in the developed framework.

**Side effects (SEs).** Drug SE information was obtained from the SIDER database. Only SEs that were MedDRA "preferred term" ("PT") were extracted and used as SE annotations for Annotated-SE dataset. The deduplicated dataset was composed of 139 516 drug-SE pairs with 4216 unique SEs (Table S6).

**Molecular representations.** A total of 29 molecular representations were investigated in this study, including 28 molecular fingerprints and InfoGraph<sup>32</sup> unsupervised learned molecular representation implemented by Torchdrug.<sup>33</sup> Detailed description of each representation can be found in Table 1. Twenty fingerprints including TT, AP, Avalon, Daylight, DFS, ASP, RDKit, PH2, PH3, MACCSFP, PubChemFP, EstateFP, KRFP, EC1024, FC1024, RAD2D, LSTAR, LingoFP, MHFP, and MAP4 were selected based on Boldini *et al.*'s study<sup>20</sup> and generated by source packages provided in the original publication. CDK-pywrapper package was used to generate all CDK fingerprints except for Daylight fingerprint. Default parameters recommended by the aforementioned packages were applied for all representations except for ECFP2048 in which the optimal parameters according to Gallo *et al.*'s study<sup>25</sup> were adopted.

### Hardware and software

All computation in this study was performed on a server with Intel® Xeon® Gold 5318Y 48-core CPU and 512 GB of RAM. Source packages provided by Boldini *et al.*<sup>20</sup> and CDK-pywrapper 0.1.1 package were implemented on Python 3.9 to generate 28 molecular fingerprints whereas Torchdrug packages was implemented on Python 3.10 to generate InfoGraph representation as described in the previous section. Other packages, including RDKit 2023.3.3, scikit-learn 1.6.1, Scipy 1.11.2, and Matplotlib 3.8.0 implemented on Python 3.11, were also used for similarity measurement, statistical analysis, and data visualization purposes. Detailed information on the version of each package used in this study can be found in .yaml files provided with the implementation scripts (see Data availability).

### Similarity metrics

In this study, the similarity between two compounds were measured by cosine similarity ( $C$ ) or Jaccard similarity ( $J$ ). Cosine similarity (eqn (1)) was applied to count, binary, and numerical representations, whereas Jaccard similarity (eqn (2)) as computed in terms of Jaccard-Needham dissimilarity by `scipy.spatial.distance.jaccard` (eqn (3)) was applied to categorical representations.

$$C(A, B) = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (1)$$

$$J(A, B) = 1 - \text{Jaccard distance}(A, B) \quad (2)$$

$$\text{Jaccard distance}(A, B) = \frac{c_{i \neq j}}{c_{i=j} + c_{i \neq j}} \quad (3)$$

where, for non-zero vectors  $A = (a_1, a_2, \dots, a_n)$  and  $B = (b_1, b_2, \dots, b_n)$ ,  $c_{ij}$  is the number of occurrences of  $A[k] = i$  and  $B[k] = j$  for  $k \leq n$ .

Similarity of categorical fingerprints was calculated in a similar manner to Boldini *et al.*'s study,<sup>20</sup> considering two bits as a match only if they possessed the exact same integer.



Table 1 List of 29 molecular representations implemented in this study

Name	Abbreviation	Implementation	Category	Type	Specified parameters	Size	Used in AgreementPred	Reference
Circular fingerprint	CircFP	CDK	Circular	Binary	—	1024	Yes	40
Local path environment fingerprint	LSTAR	jCompoundMapper	Circular	Binary	—	4096	Yes	20
Topological Molprint-like fingerprint	RAD2D	jCompoundMapper	Circular	Binary	—	4096	Yes	20
Extended connectivity fingerprint (1024 bit)	EC1024	RDKit	Circular	Binary	Radius = 2	1024	Yes	20
Extended connectivity fingerprint (2048 bit)	EC2048	RDKit	Circular	Binary	Radius = 2	2048	No	25
Functional class extended connectivity fingerprint (1024 bit)	FC1024	RDKit	Circular	Binary	Radius = 2, useFeatures = True	1024	Yes	20
Atom pair 2D fingerprint (implemented in PaDEL)	AP2DFP	CDK	Path	Binary	—	780	Yes	40
CDK fingerprint	CDKFP	CDK	Path	Binary	—	1024	No	40
Hybrid fingerprint (CDK fingerprint ignoring aromaticity)	HybridFP	CDK	Path	Binary	—	1024	Yes	40
Graph fingerprint (CDK fingerprint ignoring bond orders)	GraphFP	CDK	Path	Binary	—	1024	Yes	40
Daylight fingerprint	Daylight	CDK	Path	Binary	Depth = 7	1024	No	20
Extended CDK fingerprint (includes 25 bits for ring features and isotopic masses)	ExtFP	CDK	Path	Binary	—	1024	Yes	40
Shortest path fingerprint	SPFP	CDK	Path	Binary	—	1024	Yes	40
All shortest path fingerprint	ASP	jCompoundMapper	Path	Binary	—	4096	No	20
Depth first search fingerprint	DFS	jCompoundMapper	Path	Binary	Depth = 7	4096	Yes	20
Atom pair fingerprint	AP	RDKit	Path	Count	—	2048	Yes	20
Avalon fingerprint	Avalon	RDKit	Path	Count	—	512	Yes	20
RDKit fingerprint	RDKit	RDKit	Path	Binary	—	2048	Yes	20
Topological torsion fingerprint	TT	RDKit	Path	Count	—	2048	No	20
Pharmacophore pair fingerprint	PH2	jCompoundMapper	Pharmacophore	Binary	—	4096	No	20
Pharmacophore triplet fingerprint	PH3	jCompoundMapper	Pharmacophore	Binary	—	4096	Yes	20
LINGO fingerprint	LingoFP	CDK	String	Binary	—	1024	Yes	40
Minhashed atom pair fingerprint	MAP4	Ref.	String	Categorical	—	1024	Yes	20
Minhashed fingerprint	MHFP	Ref.	String	Categorical	—	1024	No	20
Electrotological state fingerprint	EstateFP	CDK	Substructure	Binary	—	79	Yes	40
Klekota-Roth fingerprint	KRFP	CDK	Substructure	Binary	—	4860	Yes	40
PubChem substructure fingerprint	PubChemFP	CDK	Substructure	Binary	—	881	Yes	40
Public MACCS fingerprint	MACCSFP	CDK	Substructure	Binary	—	166	Yes	40
InfoGraph graph feature	InfoGraph	Torchdrug	Unsupervised learned representation	Numerical	Learning rate (lr) = $1 \times 10^{-3}$ ; batch_size = 1024	300	Yes	32



### Single-representation similarity-based annotation prediction

Similarity-based annotation prediction was investigated for each compound in AnnoCom1000, DrugBank1000, NP1000, and Annotated-SE datasets, using 29 molecular representations. Predicted annotations for each query compound  $q$  were the union of the sets of annotations of  $N$  most similar compounds (MSCs) of the query compound, as defined in eqn (4),

$$\text{Pred}_{(r,q)} = \bigcup_{j \in \text{MSC}_r} A_j \quad (4)$$

where  $\text{Pred}_{(r,q)}$  is the predicted annotations computed using representation  $r$  for query compound  $q$ ,  $A_j$  is the set of annotations of compound  $j$ , and  $\text{MSC}_r$  is the set of  $N$  compounds from the comparison dataset with maximal similarity to query compound  $q$  as determined by the similarity metric of representation  $r$ .

For the AnnoCom1000, DrugBank1000, and NP1000 datasets, the search for MSCs was performed in batches of 50 compounds. In each batch, a query compound was compared against the remaining 9671 compounds in the Annotated-Compound dataset from which the top  $N$  most similar compounds were determined for each query compound. In contrast, the search for MSCs for Annotated-SE dataset which only contained 1376 compounds were conducted in a leave-one-out manner.

### Performance evaluation

In this study, prediction performance was evaluated using precision ( $P$ ) and recall ( $R$ ), as defined in eqn (5) and (6), respectively.

$$P = \frac{1}{C} \sum_{q=1}^C \frac{|a_q|}{|\text{Pred}_q|} \quad (5)$$

$$R = \frac{1}{C} \sum_{q=1}^C \frac{|a_q|}{|A_q|} \quad (6)$$

where  $C$  is the number of query compounds evaluated,  $A_q$  is the set of annotations of query compound  $q$ ,  $a_q$  is the set of correctly predicted annotations for query compound  $q$ , and  $\text{Pred}_q$  is the set of predicted annotations assigned to query compound  $q$ .

Prediction performance based on 1, 2, 3, 4, 5, 10, 15, 20, and 30 MSCs computed using 29 representations was compared against one another and that based on the same number of random compounds to observe the enrichment of correct annotations among top MSCs. Prediction using the same molecular representation based on  $N$  compounds of MSCs and the same number of random compounds constitute each comparing pair.

Mann-Whitney  $U$  tests were used to compare the performance each comparing pair, whereas Kruskal-Wallis tests followed by Bonferroni-corrected pairwise Mann-Whitney  $U$  tests were used to detect statistically significant difference among the performance of 29 representations.

### Similarity ranking and MSC profile of 29 representations

The similarity between three query compounds (CID = 5280343, 441764, and 446157) and the remaining 9720 compounds were ranked using 29 representations, and the average Pearson's correlation of the ranking was computed. Furthermore, the relationship between MSCs recommended by each representation and the prediction performance was also explored.

### Agreement-based data fusion

As it was hypothesized that different molecular representations captured different aspects of molecular structure and integration of structural similarity data based on multiple representations could lead to improved prediction performance, an

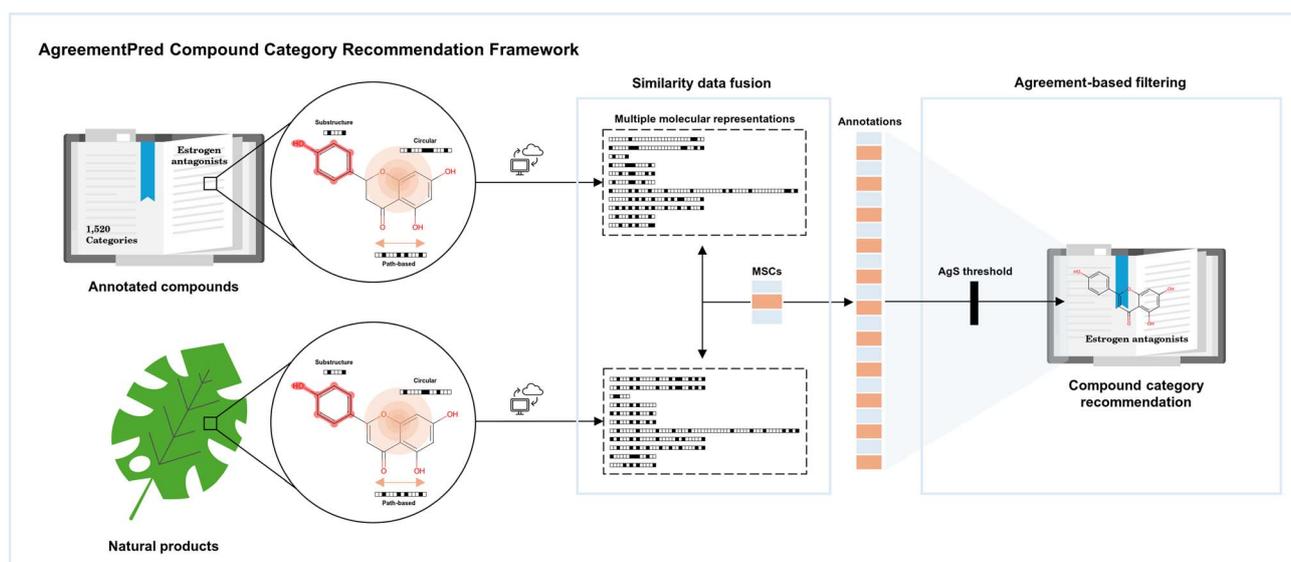


Fig. 1 Overview of AgreementPred framework. The similarity search results using 22 molecular representations were combined to improve prediction recall. Subsequently, the predicted annotations were filtered by agreement scores to enhance prediction precision.



annotation recommendation framework based on multi-representation data fusion, AgreementPred, was developed.

In AgreementPred (Fig. 1), predicted annotations resulting from multi-representation MSC-based prediction (MultiPred) of a query compound  $q$  were further filtered by agreement score (AgS), which was computed for each predicted annotations  $k$ , according to the equations below:

$$\text{MultiPred}_q = \bigcup_{r \in R} \text{Pred}_{(r,q)} \quad (7)$$

$$\text{AgS}_k = \frac{\text{count}_{\text{MultiPred}_q}(k)}{|\text{Rep}| \times N} \quad (8)$$

$$\text{AgPred}_q = \{k | k \in \text{MultiPred}_q, \text{AgS}_k > t\} \quad (9)$$

where  $\text{AgPred}_q$  is the final set of predicted annotations for query compound  $q$  using AgreementPred framework, Rep is the set of selected molecular representations incorporated in the prediction model,  $N$  is the number of MSCs used in the similarity-based prediction, and  $t$  is the predefined threshold of AgS used for the prediction model.

The prediction performance of AgreementPred was evaluated on AnnoCom1000, DrugBank1000, NP1000, and Annotated-SE datasets, comparing different  $t$  and  $N$  parameters.

### Method comparison

To benchmark AgreementPred, its performance in annotation prediction was tested using PubChem annotations (AnnoCom1000), second-, and fourth-level ATC annotations, comparing with two previously published ATC-predicting models, SD-ATC<sup>11</sup> and iSEA,<sup>24</sup> as well as EC1024 similarity-based prediction.

For reasons mentioned in the Introduction section, only methods which adopted molecular structure as the sole input were considered for comparison. SD-ATC employed KRFP as the molecular representation and utilized network-based inference approach to extract the relationship between molecular substructures and ATC classes, whereas iSEA utilized similarity ensemble approach using the average similarity of 3 molecular representations (CDKFP, PubChemFP, and MACCSFP) to quantify the relation of a given drug to each ATC class based on the level of molecular similarity between the drug and drug set belonging to each class.

SuperPred frameworks<sup>25,34,35</sup> also adopted molecular structure as the sole input of the models. However, extensive pre-processing of training data was required for SuperPred approaches, especially SuperPred3.0 in which single-label training dataset was mandatory for logistic regression model. Therefore, SuperPred frameworks were not selected to be compared in this section.

The benchmark datasets for second- and fourth-level ATC used in this study were derived from the training set containing 1151 approved drugs provided in iSEA original publication. A subset containing 1107 compounds with obtainable PubChem CIDs and PubChem's canonical SMILES was used in this study. Second-level ATC labels were obtained from the original

dataset, whereas fourth-level ATC labels were extracted from the DrugBank database. The ATC datasets were divided into 22 batches containing 50–51 compounds. AgreementPred and SD-ATC were implemented using the same batches of testing data for all datasets.

As iSEA required computing average similarity based on 3 molecular representations with 1000 permutations for every drug-ATC pair, the framework was presumed to be infeasible for a dataset with a large number of classes such as PubChem annotations and fourth-level ATC. Therefore, iSEA was compared with other methods only for the performance on second-level ATC prediction, and the results were directly derived from the original publication without implementation.

### Application

AgreementPred was applied on 321 605 unannotated compounds from drug and natural product databases. After eliminating redundant representations, 22 molecular representations, namely CircFP, LSTAR, RAD2D, EC1024, FC1024, AP2DFP, HybridFP, GraphFP, ExtFP, SPFP, DFS, AP, Avalon, RDKit, PH3, LingoFP, MAP4, EstateFP, KRFP, PubChemFP, MACCSFP, and InfoGraph, were incorporated (summarized in Table 1), using MSC and agreement score threshold of 1 and 0.1, respectively. The predicted annotations were assigned to each query compound, and the results were further analyzed for plausibility.

## Results

### Single-representation similarity-based annotation prediction

AnnoCom1000, DrugBank1000, and NP1000 datasets showed a similar pattern of performance resulting from 29 molecular representations (Fig. 2). The performance of MSC-based prediction (Fig. 2A, C and E) was significantly higher ( $p$ -value  $< 0.05$ ) than that of random prediction (Fig. 2B, D and F) for all comparing pairs except for PH2 at various MSCs. Significant difference (Kruskal–Wallis  $p$ -value  $< 0.05$ ) in recall and precision was detected among MSC-based prediction of 29 representations at every MSC, while no difference was detected for prediction performance among 29 representations based on random compounds. However, post-hoc Mann–Whitney  $U$  tests indicated comparable performance among most representations (Bonferroni-corrected  $p$ -value  $> 0.05$ ), except for PH2, PH3, EStateFP, AP2DFP, and GraphFP in which the recall and precision were significantly lower than those of other representations.

The results suggested that similarity-based prediction was somewhat effective for category annotations, comprising ATC and MeSH classification, as the annotations significantly enriched among compounds with high similarity to query compounds, aligning with a well-established concept that chemical compounds with a similar structure tend to possess similar properties.<sup>36</sup> The results were also consistent with the findings of Boldini *et al.*'s which showed that while different molecular fingerprints performed best on different datasets, pharmacophore-based fingerprints tended to underperform other types.<sup>20</sup>



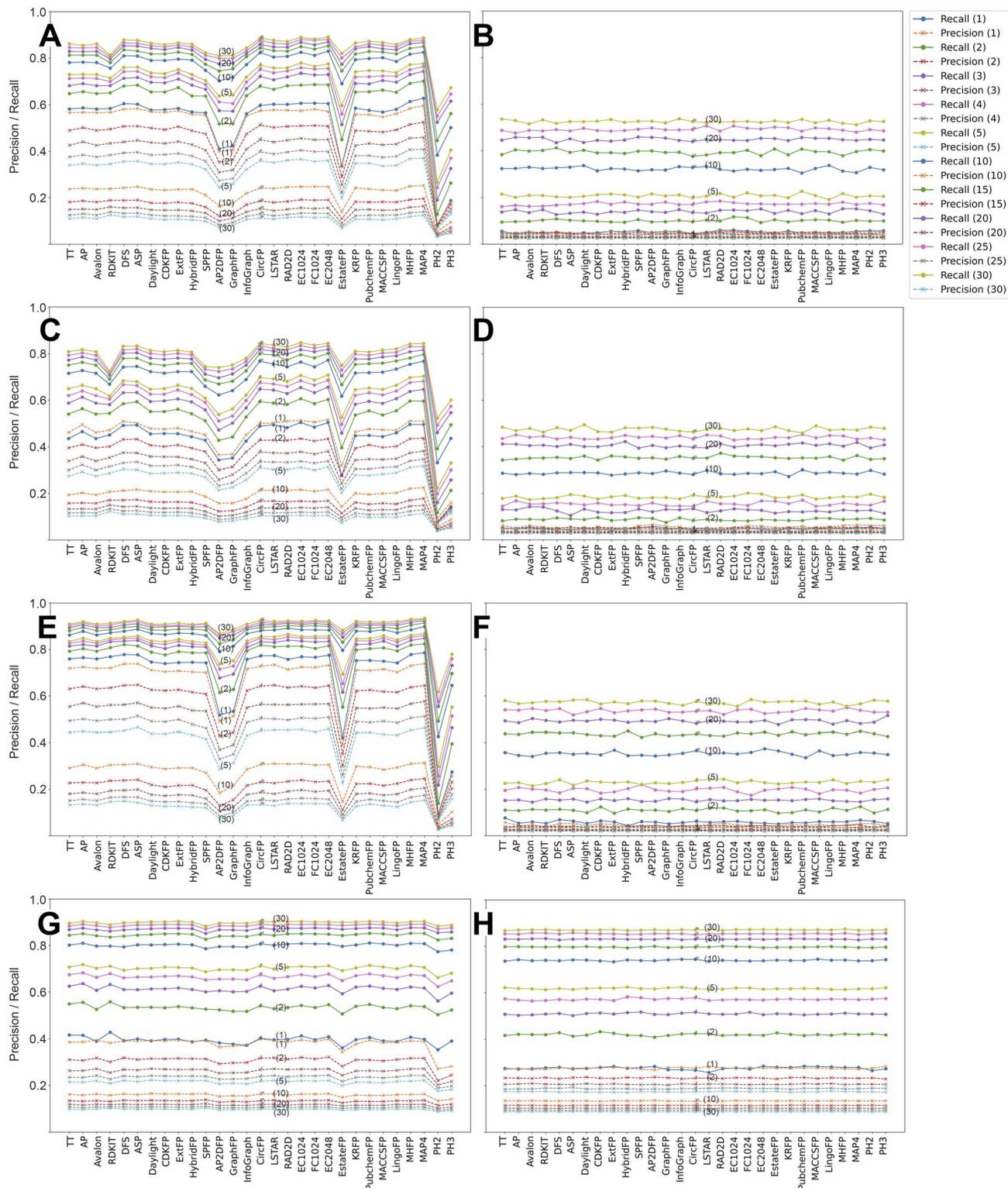


Fig. 2 Single-representation similarity-based annotation prediction. Prediction recall (in solid line) and precision (in dashed line) of similarity-based prediction based on MSCs (left column) and random compounds (right column) computed using 29 molecular representations on AnnoCom1000 (A and B), DrugBank1000 (C and D), NP1000 (E and F), and Annotated-SE (G and H) datasets. Bracketed numbers in the legend show the number of MSCs or random compounds used for prediction.

Comparing the performance of similarity-based prediction on drug and natural product datasets, it was discovered that the overall recall and precision were significantly higher ( $p$ -value <

0.05) for NP1000 than the DrugBank1000 dataset (Fig. 2C and E), except for the recall of PH2, and the precision of PH3, EStateFP, AP2DFP, GraphFP, and SPFP at various MSCs. The



difference possibly stemmed from a higher number of annotations (971 vs. 544) and compound-specific annotations (339 vs. 234) in DrugBank1000 than in NP1000, indicating that the performance of similarity-based prediction could be compromised by the diversity of annotations. This problem could be mitigated by annotation screening and/or grouping; however, elimination or manipulation of labels might also lead to loss of relevant information.

For the Annotated-SE dataset, the difference between the comparing pairs of MSCs and random compounds were not as noticeable as in AnnoCom1000, DrugBank1000, and NP1000 (Fig. 2G and H), however, Mann-Whitney *U* tests resulted in a *p*-value lower than 0.05 for all comparing pairs, except for the precision of PH2, PH3, and EStateFP at various MSCs.

It was noteworthy that the average number of annotations per compound were 5.62 vs. 101.78, and the maximum number of annotations per compound were 47 (dexamethasone) vs. 742

(pregabalin) in Annotated-Compound and Annotated-SE dataset, respectively. In particular, high occurrences of some SEs, such as headache, nausea, and vomiting, were likely to be responsible for high apparent performance of prediction based on random compounds. Nevertheless, MSC-based predictions showed significant difference in recall and precision resulting from 29 representations (Kruskal-Wallis *p*-value < 0.05) at every MSC, while no difference was shown among random predictions. The pattern of performance of 29 molecular representations also differed from that on Annotated-Compound datasets, with RDKit fingerprint obtaining prominent recall especially at 1 MSC, and only PH2, PH3, and EStateFP showed notably inferior performance to other representations.

The results suggested that molecular similarity might be insufficient to deliver a reliable SE prediction based on currently available data. Unlike pharmacological categories which are established based on experimental results, drug SEs are

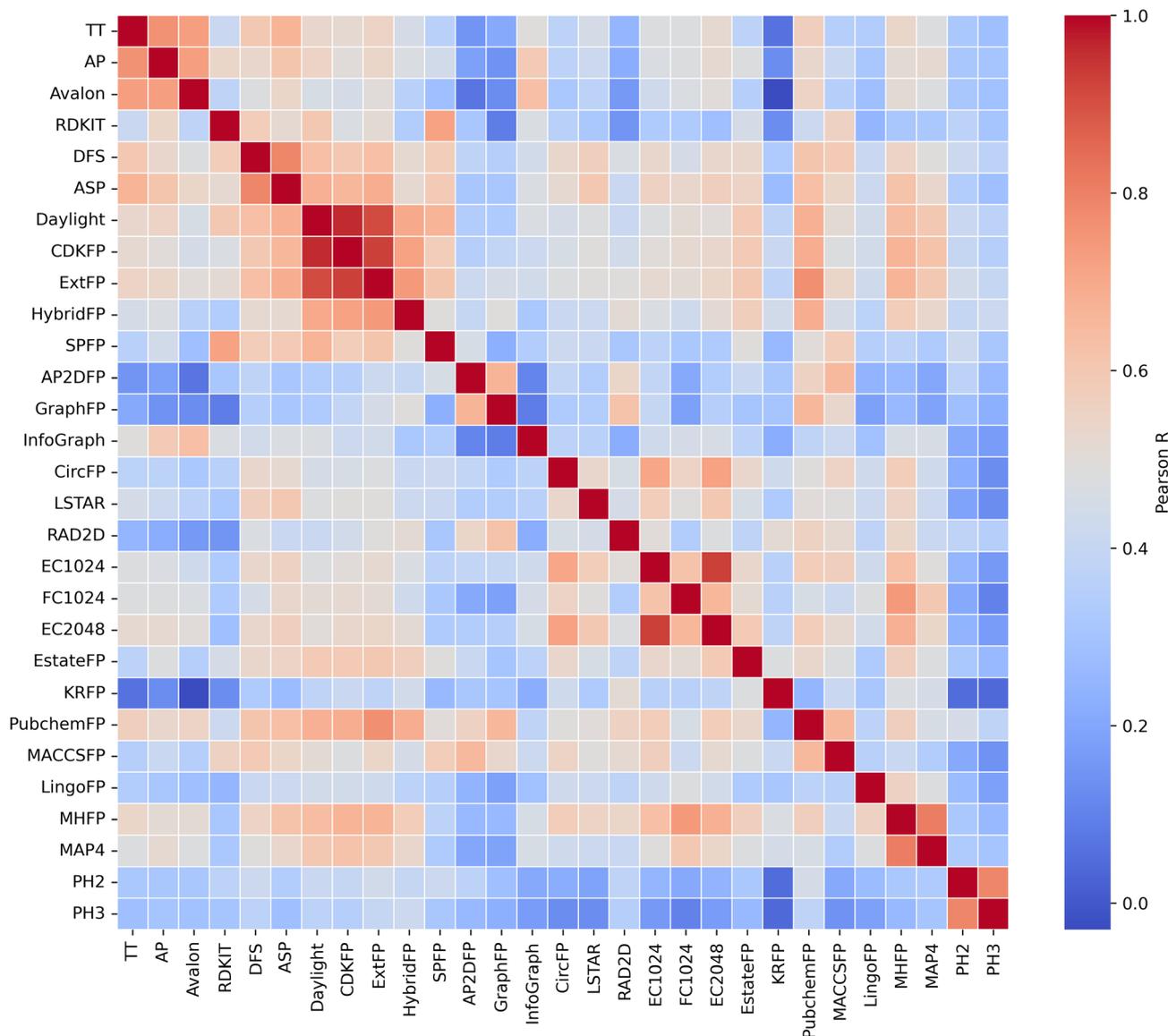


Fig. 3 Pearson's correlation among the similarity ranking profile of 29 molecular representations.



typically defined based on observation during randomized controlled clinical trials. Consequently, the SEs of each drug vary significantly in frequency, severity, and clinical relevance, adding considerable complexity to the prediction task that may necessitate more sophisticated approaches.

### Similarity ranking profile of 29 representations

Fig. 3 shows Pearson's correlation among 29 molecular representations. High correlation, indicating similar ranking profile, was observed between representations generated by similar computing algorithms, such as EC1024 and EC2048, PH2 and

PH3, MHFP and MAP4. Considering the prediction performance of each representation (see previous section), it was noteworthy that representations with different ranking profiles, as indicated by low correlation, resulted in comparable prediction performance. This suggested that different aspects of similarity might be responsible for the retrieval of different annotations, as exemplified by the similarity-based prediction of an antihypertensive drug, diltiazem, whose annotations include cardiovascular system, cardiovascular agents, antihypertensive agent, membrane transport modulators, and vasodilator agents.

MSCs of diltiazem computed using 29 representations are shown in Fig. 4A with the corresponding structures and representations shown in Table 2. Of 22 compounds, 3 compounds possessed the annotations of diltiazem: antihypertensive agent, cardiovascular agents, and cardiovascular system were among benazepril's annotations, predicted by SPFP and Avalon; cardiovascular agents and vasodilator agents belonged to tadalafil, predicted by AP; while cardiovascular agents and membrane transport modulators were retrieved by EStateFP among the annotations of cocaine. It is also worth mentioning that annotations relating to cardiovascular system were common among the annotations of different compounds, predicted by different molecular representations.

Consequently, it was further hypothesized that integration of multiple representations in similarity-based prediction might lead to improved performance relative to single-representation prediction.

## Agreement-based data fusion

Although similarity-based prediction resulted in moderately satisfying performance, there remained two key disadvantages. Firstly, the recall and precision of the prediction greatly diverged. As shown in Fig. 2, while recall increased with the number of MSCs, precision sharply decreased. Secondly, since there was no guarantee that structurally similar compounds would be present in the annotated dataset, MSC-based prediction without a similarity threshold could lead to poor performance due to the absence of compounds truly similar to the query compound. On the other hand, setting a strict similarity threshold might cause certain annotations to be missed, particularly when compounds within certain annotation groups share similar substructures but differ in other parts that lower the overall similarity score. Determining an optimal threshold could present an additional challenge.

AgreementPred (Fig. 1), a category recommendation framework for drugs and natural products based on multi-representation data fusion, was developed to address these problems. The framework was devised based on the hypothesis that the degree of agreement among different molecular representations in identifying MSCs of a query compound could indicate the overall similarity of the pair of compounds, and the overall similarity could, in turn, indicate the degree of certainty the pair belongs to the same categories. Moreover, annotations that are common among different MSCs predicted using

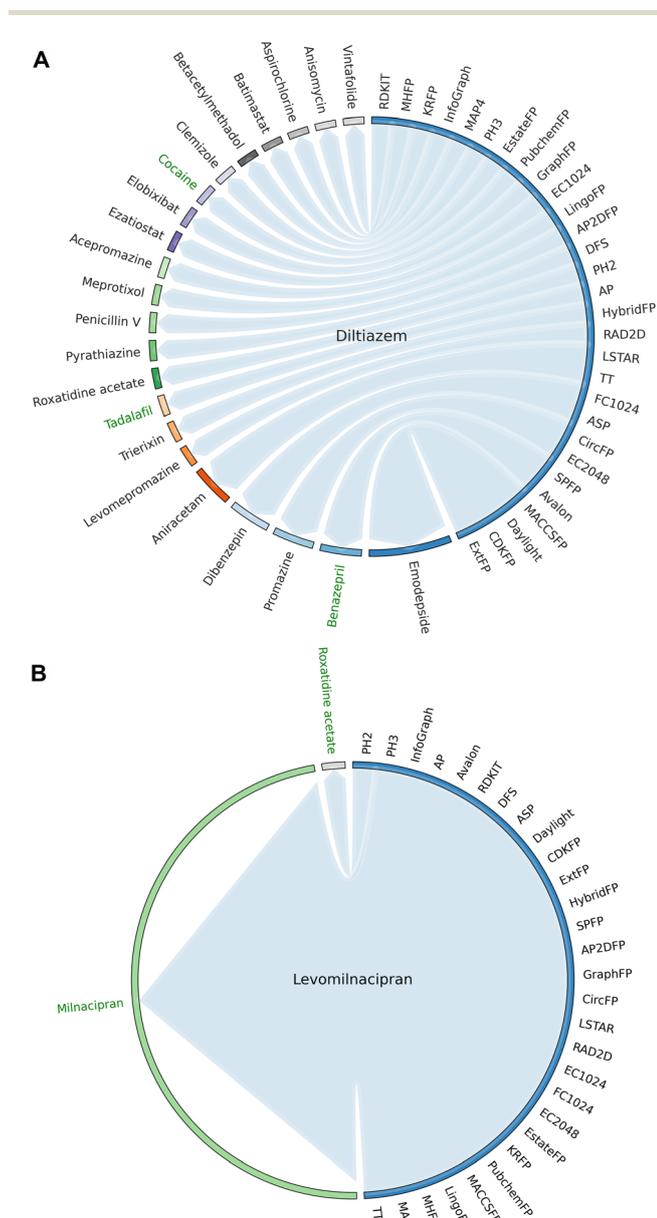
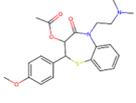
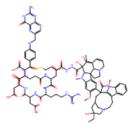
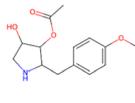
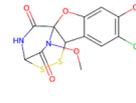
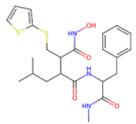
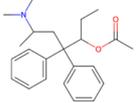
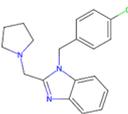
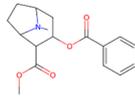
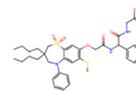
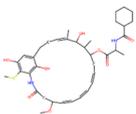
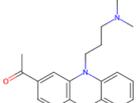
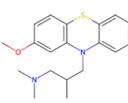
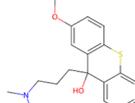
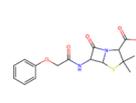
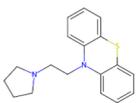
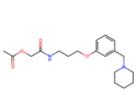
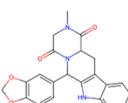
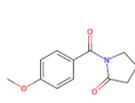
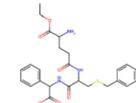
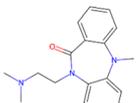
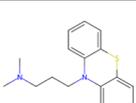
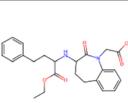
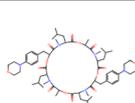


Fig. 4 The degree of agreement indicates overall similarity. The MSCs of diltiazem (A) and levomilnacipran (B) identified through similarity search using 29 molecular representations. Compounds that possessed one or more annotations of the query compound are shown in green.



Table 2 Molecular structure of MSC of diltiazem identified through similarity search using 29 molecular representations

Query compound 	RDKit 	MHFP 	KRFP 	InfoGraph 
MAP4 	PH3 	EStateFP 	PubChemFP 	HybridFP 
ECFP1024 	RAD2D 	LingoFP 	AP2DFP 	DFS 
PH2 	AP 	LSTAR, TT 	GraphFP 	ASP, FCFP1024 
CircFP, EC2048 	SPFP, Avalon 	ExtFP, Daylight, CDKFP, MACCSFP 		

different molecular representations were also more likely to be related to the query compound.

This hypothesis was supported by diltiazem's annotation prediction (see previous section) and the MSC profile computed

by 29 molecular representations of levomilnacipran, in comparison to previously mentioned diltiazem. Whereas 29 representations identified 22 different compounds as the MSC of diltiazem, 28 out of 29 representations identified

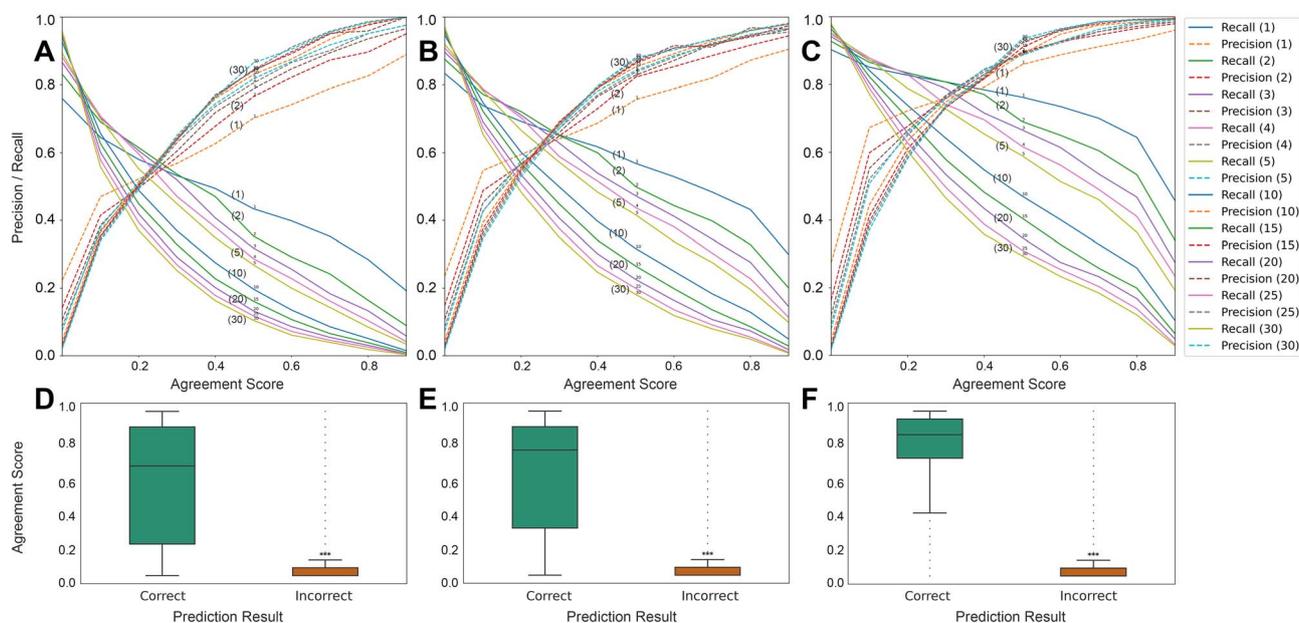


Fig. 5 Prediction performance of AgreementPred framework. Recall (in solid line) and precision (in dashed line) of AgreementPred framework adopting various  $N$  of MSCs and the threshold ( $t$ ) of agreement score on DrugBank1000 (A), AnnoCom1000 (B), and NP1000 (C) dataset; and agreement score comparison between correct and incorrect prediction of DrugBank1000 (D), AnnoCom1000 (E), and NP1000 (F) dataset ( $N = 1$ ). Bracketed numbers in the legend show  $N$  of MSCs used for prediction.



milnacipran, a stereoisomer of levomilnacipran as the most similar compound of levomilnacipran. Milnacipran possessed 10 out of 12 of levomilnacipran's annotations, while benazepril possessed 4 out of 14 of diltiazem's annotations, reflecting that prediction performance increased with degree of agreement.

Leveraging this finding, 22 representations, 1 from each group of representations that were within the same category and were highly correlated (Pearson's correlation > 0.75), as shown in Fig. 3, were incorporated into AgreementPred framework. Seven fingerprints, including TT, ASP, Daylight, CDKFP, EC2048, MHFP, and PH2, were excluded to prevent biased agreement. Annotations predicted by the 22 representations were subsequently filtered by a preset threshold of agreement score which was computed for each of the predicted annotations as the indicator of the degree of agreement (eqn (8)). In this way, prediction recall could be improved through the pooling of predicted annotations resulting from multiple representations, and prediction precision could be enhanced by agreement-based filtering, in which only the annotations of a compound with high overall similarity to the query compound or the annotations shared among multiple MSCs would be predicted for the query compound.

The performance of AgreementPred on 3 sample datasets adopting various  $N$  of MSCs and the threshold ( $t$ ) of agreement score is shown in Fig. 5. At  $t = 0$ , the performance of AgreementPred was comparable to the performance of similarity-based prediction using equivalent number of compounds. However, unlike similarity-based prediction, recall and precision of AgreementPred demonstrated convergence with increasing  $t$  of agreement score up to a certain point, where precision began to outweigh recall. Thus, by adjusting  $N$  and  $t$ , the preferred balance of recall and precision could be achieved.

Moreover, as shown in Fig. 5D–F, the agreement score of correct prediction was significantly higher (Mann–Whitney  $U$   $p$ -value <  $10^{-30}$ ) than that of incorrect prediction in all datasets, confirming the correlation between agreement score and prediction accuracy. Hence, in AgreementPred, predicted annotations could be sorted by their agreement scores as the indicators of prediction confidence.

## Method comparison

AgreementPred showed superiority in the balance of prediction recall and precision to other models in all comparison tasks, including PubChem annotations, second-, and forth-level ATC prediction. As shown in Table 3, at MSC and AgS threshold of 2 and 0.0, respectively, the resulting precision of AgreementPred was comparable to that of iSEA and SD-ATC in all tasks while the recall was notably higher. At MSC and AgS threshold of 1 and 0.1, respectively, AgreementPred showed inferior recall to iSEA in second-level ATC task, and comparable recall to SD-ATC in second- and fourth-level ATC tasks, however, the precision was significantly superior.

On the PubChem annotation prediction task (AnnoCom1000), the performance of SD-ATC was shown to be greatly inferior to EC1024 and AgreementPred (Fig. 6). This possibly stemmed from the task-specificity of SD-ATC which was

**Table 3** Prediction performance of AgreementPred in comparison to iSEA, SD-ATC, and EC1024 similarity-based prediction on second-, fourth-level ATC, and AnnoCom1000 datasets<sup>a</sup>

	AgreementPred (MSC = 1; AgS > 0.0)	AgreementPred (MSC = 1; AgS > 0.1)	AgreementPred (MSC = 2; AgS > 0.0)	AgreementPred (MSC = 2; AgS > 0.1)	AgreementPred (MSC = 2; AgS > 0.1)	EC1024 (MSC = 1)	EC1024 (MSC = 2)	iSEA ( $L = 10$ )	SD-ATC ( $L = 10$ )
<b>Second-level ATC</b>									
Recall <sup>b</sup>	0.745 (1041/1397)	0.633 (884/1397)	0.633 (884/1397)	0.670 (937/1397)	0.670 (937/1397)	0.523 (731/1397)	0.601 (840/1397)	0.748 (1128/1509)	0.671 (937/1397)
Precision <sup>b</sup>	0.139 (1041/7481)	0.363 (884/2436)	0.090 (1144/12867)	0.322 (937/2914)	0.322 (937/2914)	<b>0.491 (731/1488)</b>	0.363 (840/2315)	0.098 (1128/11510)	0.085 (937/11070)
<b>Fourth-level ATC</b>									
Recall	0.579	0.480	<b>0.635</b>	0.529	0.529	0.369	0.459	—	0.478
Precision	0.158	0.348	0.091	0.311	0.311	<b>0.365</b>	0.307	—	0.060
<b>AnnoCom1000</b>									
Recall	0.833	0.739	<b>0.875</b>	0.772	0.772	0.607	0.685	—	0.374
Precision	0.236	0.547	0.148	0.487	0.487	<b>0.574</b>	0.510	—	0.173

<sup>a</sup>  $L$ : prediction length. <sup>b</sup> Recall and precision in second-level ATC prediction task are computed in the same manner as iSEA,<sup>24</sup> by dividing the total number of correct predictions by the total number of labeled classes (recall); and by the total number of predictions (precision), respectively, as specified in the brackets.



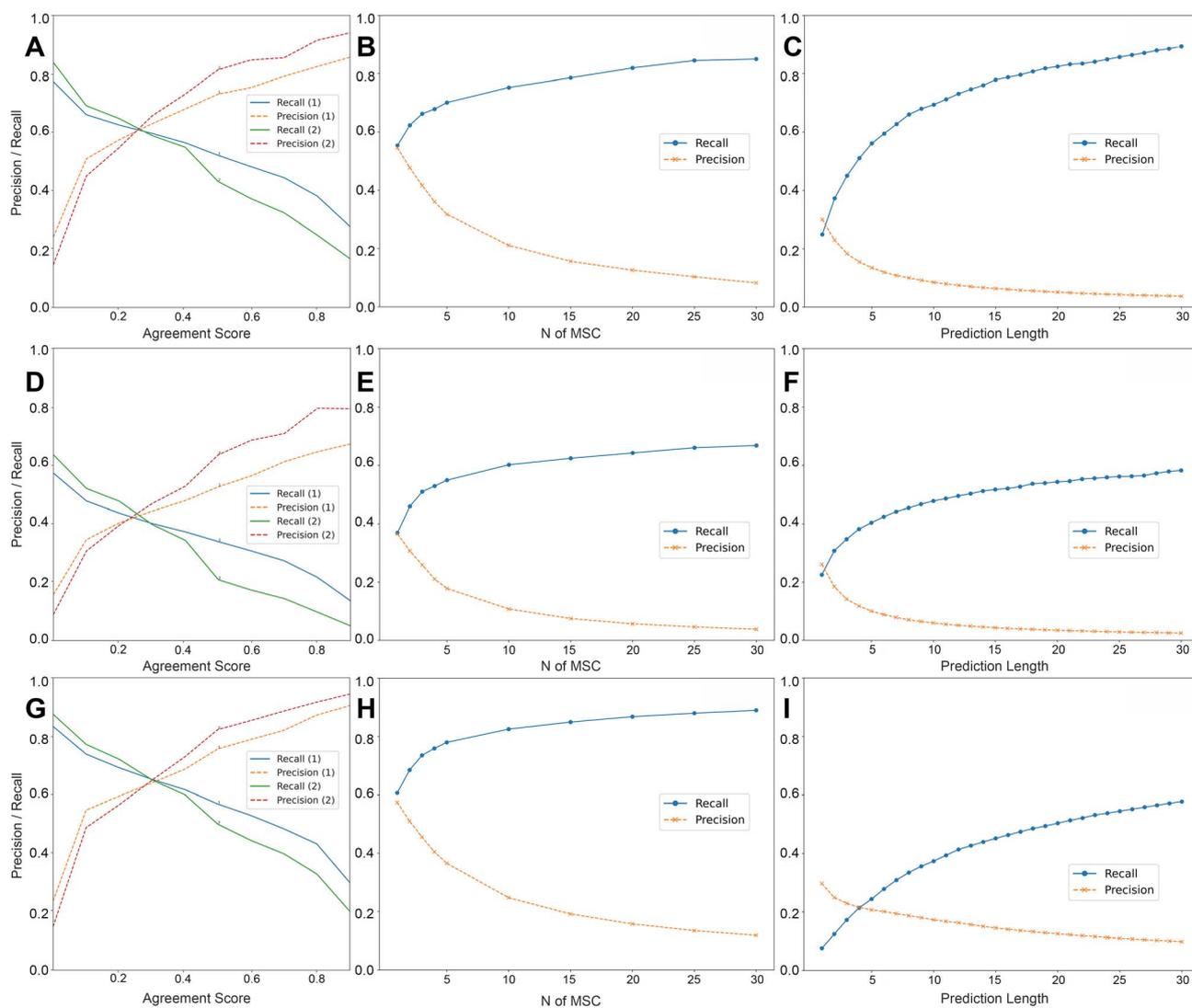


Fig. 6 Method comparison. Prediction recall (in solid line) and precision (in dashed line) of AgreementPred (left column), EC1024 similarity-based prediction (center column), and SD-ATC (right column) on second-level ATC (A–C), fourth-level ATC (D–F), and AnnoCom1000 (G–I) dataset. Bracketed numbers in the legend show  $N$  of MSCs used for prediction in AgreementPred.

Table 4 Annotations predicted by AgreementPred and the corresponding supporting literature for apigenin, licochalcone C, and phillyrin

Compound name	CID	Prediction	Agreement score	Supporting literature
Apigenin	5280443	Protective agent	0.41	41 and 42
		Hormone antagonist	0.41	43–45
		Anticarcinogenic agents	0.18	42, 46 and 47
		Tyrosine kinase inhibitor	0.18	41 and 47
		Angiogenesis inhibitor	0.18	48–50
		Prostaglandin antagonists	0.14	51–53
Licochalcone C	9840805	Anti-inflammatory agents	0.14	41, 42 and 52
		Antineoplastic agent	0.64	54 and 55
		Angiogenesis inhibitor	0.36	54
		Growth inhibitors	0.36	55
Phillyrin	101712	Antihypertensive agent	0.32	56 and 57
		Hypolipidemic agents	0.32	58 and 59
		Anti-inflammatory agents	0.23	60–63
		Cyclooxygenase inhibitor	0.18	62



optimized for ATC prediction and inherent difference between the two tasks. In this regard, the prediction performance of SD-ATC and AgreementPred on each PubChem annotation was further explored. Detailed comparison of prediction precision of each annotation by the two methods is shown in Table S7. It was demonstrated that SD-ATC, utilizing a network-based inference approach, suffered greatly from class imbalance in the PubChem annotation dataset, and clearly biased toward annotations with high occurrence. For example, SD-ATC predicted 'enzyme inhibitor', which was the annotation with the highest occurrence, for 997 compounds out of 1000 compounds in AnnoCom1000 dataset. As a result, SD-ATC was only able to correctly predict 86 out of 872 unique annotations with prediction length of 10 (10 000 predictions in total). In contrast, AgreementPred was shown to be more tolerant of a highly diverse and imbalanced dataset. It correctly predicted 665 out of 872 unique annotations among 9403 predictions in total. Mean precision across all annotations for SD-ATC and AgreementPred were 0.06 and 0.41, respectively.

Extended connectivity fingerprint (ECFP) is widely accepted for its superior performance in bioactivity prediction to other molecular fingerprints.<sup>19</sup> However, similarity-based prediction using ECFP as molecular representation implemented in this study revealed comparable performance to most other molecular representations (see Single-representation similarity-based annotation prediction section). Therefore, EC1024 was employed here as the representative single-representation prediction method.

As shown in Fig. 6, EC1024 and SD-ATC exhibited a pattern in which recall and precision continued to diverge as the number of MSCs or prediction length increased, until eventually reaching a plateau. For both methods, precision peaked at small values of MSCs or shorter prediction length, but this improvement was achieved at the expense of the reduced recall. Notably, EC1024 similarity-based prediction with 2–5 MSCs achieved a balance of recall and precision only slightly inferior to that of AgreementPred. However, owing to the use of agreement scores, AgreementPred demonstrated distinct advantages including greater adjustability, presence of prediction filtering and a confidence indicator. These features are critical, as they help to mitigate poor prediction performance that may arise in single-representation similarity-based methods when annotated compounds with sufficient similarity to the query are not present in the dataset. Moreover, by applying higher agreement score thresholds, AgreementPred could achieve substantially higher precision, further underscoring its superiority over single-representation approaches.

## Application

AgreementPred was applied to predict categories of 321 605 unannotated compounds from drug and natural product databases, using 22 selected molecular representations (Table 1). Before agreement-based filtering, 12 691 685 category labels were recommended for 321 605 compounds. Subsequently, 9 802 758 predictions were removed using an agreement score threshold of 0.1, as described in the Material and methods,

giving a total of 2 888 927 predicted category labels for 321 596 compounds (Table S8). After the concatenation of the Annotated-Compound dataset with the final prediction result, 2 943 602 category labels were provided for 331 317 compounds (Table S9). The average number of category labels per compound in the final concatenated dataset was  $8.9 \pm 5.0$ , increasing from that in the original Annotated-Compound dataset ( $5.6 \pm 4.2$ ).

Predictions were analyzed for a subset of relatively well-studied compounds that remained unannotated in PubChem database, namely apigenin, licochalcone C, and phillyrin. These compounds have been extensively investigated in previous pharmacological studies, providing a valuable reference for external validation. The predicted categories for these compounds were all derived from annotated compounds with high structural similarity. Mean similarity values across MSCs resulting from 22 molecular representations of the three compounds were 0.89, 0.83, and 0.77, respectively, indicating high plausibility of the prediction. Indeed, Table 4 showed that the key pharmacological effects predicted for each compound were consistent with findings reported in previously published literature.

Furthermore, in an attempt to relate the pharmacological categories of chemical components to the pharmacological properties of medicinal herbs for further mechanistic study of herbal medicines as mentioned in the Introduction section, the resulting annotations of natural products contained in 3 prominent traditional Chinese medicine (TCM) herbs, Ephedrae Herba (Mahuang), Rhei Radix et Rhizoma (Dahuang), and Salviae Miltiorrhizae Radix et Rhizoma (Danshen), were investigated. It was discovered that the pharmacological categories widely recognized as the main pharmacological properties of all 3 herbs were among top 20 annotations of highest occurrences.

In detail, Mahuang, a TCM herb well-recognized for its effects on respiratory and cardiovascular systems,<sup>37</sup> comprised 42 and 39 compounds in 'cardiovascular agents' and 'respiratory system agents' categories, ranking top 16 and 18 of annotations with the highest occurrences, respectively. Among these, 35 and 34 compounds were predicted by AgreementPred.

In Dahuang, an herb well-renowned for its strong laxative effect,<sup>38</sup> 74 compounds in total were predicted to possess pharmacological categories 'laxative' and 'cathartics', ranking top 9 and 10, respectively. Lastly, in Dashen, an herb well-recognized for its uses in various cardiovascular diseases,<sup>39</sup> hematologic agents' and 'anticoagulants' were predicted for 150 and 112 compounds, ranking top 9 and 11, respectively.

These results tentatively lent empirical support to AgreementPred's predictive capability and revealed an inherent relationship between the pharmacological properties of herbs and the pharmacological categories of their constituents, offering valuable insights into further mechanistic studies of herbal medicines.

## Discussion

In this study, AgreementPred, a simple and completely interpretable category recommendation framework for drugs and



natural products was proposed. Unlike machine-learning approaches that require a large amount of training data, AgreementPred only requires a few similar compounds in the annotated dataset for reasonable predictions. As such, the framework also possessed high tolerance of class imbalance compared to network-based approach, in which the occurrences of predicted annotations were directly proportional to the occurrences of the annotations in the dataset. Moreover, for AgreementPred, each predicted annotation can be transparently traced back to the specific annotated compounds that contributed to the prediction, allowing the rationale behind each annotation to be evaluated, serving as another significant advantage over other sophisticated approaches.

Nevertheless, the proposed framework is far from perfect. Its main limitation lies in its inability to “think outside the box”. Unlike machine learning or network-based approaches that are capable of recognizing latent, complex patterns across high-dimensional data spaces and uncovering non-obvious association, the framework is inherently constrained by its reliance on known and explicitly defined similarity. As a result, this framework is not capable of identifying compound-specific properties or a novel class of bioactivities that are not shared by structurally similar compounds.

To mitigate this limitation, additional approaches could be integrated into the framework. For example, alternative molecular representations, such as physicochemical property profiles or knowledge graph embeddings, could be utilized to provide complementary aspects of a compound beyond its chemical structure. Moreover, natural language processing techniques and large language models could be employed to explore semantic relationships among annotation terms, thereby enabling the extraction of related annotations even when the compounds do not possess sufficient structural similarity. Collectively, these strategies have the potential to improve the framework's generalizability while alleviating the trade-off between interpretability and discovery power.

## Conclusion

In the proposed framework, AgreementPred, categories of drugs and natural products were predicted through multi-representation structural similarity data fusion and subsequently subjected to agreement-based filtering. The prediction performance of the framework was validated on ATC and PubChem annotation datasets and was shown to be superior in terms of recall-precision balance to existing equivalent methods. It also offers significant advantages over existing approaches in explainability, adjustability, and tolerance to limited data points and class imbalance. However, the framework suffers from inability to predict properties that are not shared by structurally similar compounds.

AgreementPred was applied to predict categories of 321 605 unannotated compounds from drug and natural product databases. A total of 2 888 927 categories were recommended for 321 596 compounds. The results provided preliminary support for the framework's predictive capability, reasonably annotated pharmacological categories for numerous natural products, and

outlined a relationship between the pharmacological effects of herbs and their components, offering potential insights into drug discovery and future mechanistic study of herbal medicines.

## Conflicts of interest

The authors declare that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

## Abbreviations

ATC	Anatomical therapeutic chemical
MeSH	Medical subject headings
SE	Side effect
FP	Fingerprint
ECFP	Extended connectivity fingerprint
AP	Atom pair fingerprint
PHFP	Pharmacophore fingerprint
PH2	Pharmacophore pair fingerprint
PH3	Pharmacophore triplet fingerprint
EStateFP	Electrotopological state fingerprint
AP2DFP	Atom pair 2D fingerprint
GraphFP	Graph fingerprint
SPFP	Shortest path fingerprint
AgS	Agreement score
MSC	Most similar compound

## Data availability

The data, scripts and instruction necessary to implement AgreementPred, and to reproduce the key results presented in this study are available on GitHub (<https://github.com/ChayanisSu/AgreementPred>) and Zenodo (<https://zenodo.org/records/17169919>) repositories.

Supplementary information is available. See DOI: <https://doi.org/10.1039/d5dd00329f>.

## Acknowledgements

This work was supported by research grants from the National Natural Science Foundation of China (Grant No. T2341008 and 82305047), the Innovation Team and Talents Cultivation Program of National Administration of Traditional Chinese Medicine (Grant No. ZYYCXTD-D-202405), and the Pilot Project for Disciplinary Breakthroughs of Ministry of Education ('Prevention and Treatment of Multi-System Comorbid Diseases with Traditional Chinese Medicine').

## References

- 1 A. G. Atanasov, *et al.*, Discovery and resupply of pharmacologically active plant-derived natural products: A review, *Biotechnol. Adv.*, 2015, **33**(8), 1582–1614.



- 2 M. R. Lee, The history of Ephedra (ma-huang), *J R Coll Physicians Edinb*, 2011, **41**(1), 78–84.
- 3 X. Z. Su and L. H. Miller, The discovery of artemisinin and the Nobel Prize in Physiology or Medicine, *Sci. China: Life Sci.*, 2015, **58**(11), 1175–1179.
- 4 N. B. Cech and N. H. Oberlies, From plant to cancer drug: lessons learned from the discovery of taxol, *Nat. Prod. Rep.*, 2023, **40**(7), 1153–1157.
- 5 P. Zhang, *et al.*, Network pharmacology: towards the artificial intelligence-based precision traditional Chinese medicine, *Briefings Bioinf.*, 2023, **25**(1), bbad518.
- 6 W. Zhou, *et al.*, FordNet: Recommending traditional Chinese medicine formula via deep neural network integrating phenotype and molecule, *Pharmacol. Res.*, 2021, **173**, 105752.
- 7 S. Li, Mapping ancient remedies: applying a network approach to traditional Chinese medicine, *Science*, 2015, **350**(6262 Suppl), S72–S74.
- 8 M. W. Mullowney, *et al.*, Artificial intelligence for natural product drug discovery, *Nat. Rev. Drug Discovery*, 2023, **22**(11), 895–916.
- 9 WHO, Anatomical Therapeutic Chemical (ATC) Classification, 2025, Available from: <https://www.who.int/tools/atc-ddd-toolkit/atc-classification>.
- 10 P. Das and D. H. Mazumder, A Comprehensive Survey of Studies on Predicting Anatomical Therapeutic Chemical Classes of Drugs, *ACM Comput. Surv.*, 2024, **57**(3), 1–31.
- 11 Y. Peng, *et al.*, Drug repositioning by prediction of drug's anatomical therapeutic chemical code via network-based inference approaches, *Briefings Bioinf.*, 2021, **22**(2), 2058–2072.
- 12 P. N. Hameed, *et al.*, A two-tiered unsupervised clustering approach for drug repositioning through heterogeneous data integration, *BMC Bioinf.*, 2018, **19**(1), 129.
- 13 N.L.o. Medicine, Medical Subject Headings (MeSH), 2025, Available from: <https://www.nlm.nih.gov/mesh/meshhome.html>.
- 14 F. S. Chen and Z. R. Jiang, Prediction of drug's Anatomical Therapeutic Chemical (ATC) code by integrating drug-domain network, *J. Biomed. Inf.*, 2015, **58**, 80–88.
- 15 Z. Liu, *et al.*, Similarity-based prediction for Anatomical Therapeutic Chemical classification of drugs by integrating multiple data sources, *Bioinformatics*, 2015, **31**(11), 1788–1795.
- 16 L. Chen, J. Xu and Y. Zhou, PDATC-NCPMKL: Predicting drug's Anatomical Therapeutic Chemical (ATC) codes based on network consistency projection and multiple kernel learning, *Comput. Biol. Med.*, 2024, **169**, 107862.
- 17 T. Olson and R. Singh, Predicting anatomic therapeutic chemical classification codes using tiered learning, *BMC Bioinf.*, 2017, **18**(8), 266.
- 18 H. Zhao, *et al.*, RNPredATC: A Deep Residual Learning-Based Model With Applications to the Prediction of Drug-ATC Code Association, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2023, **20**(5), 2712–2723.
- 19 L. David, *et al.*, Molecular representations in AI-driven drug discovery: a review and practical guide, *J. Cheminf.*, 2020, **12**(1), 56.
- 20 D. Boldini, *et al.*, Effectiveness of molecular fingerprints for exploring the chemical space of natural products, *J. Cheminf.*, 2024, **16**(1), 35.
- 21 G. Liu, *et al.*, Deep learning-guided discovery of an antibiotic targeting *Acinetobacter baumannii*, *Nat. Chem. Biol.*, 2023, **19**(11), 1342–1350.
- 22 K. Yang, *et al.*, Analyzing Learned Molecular Representations for Property Prediction, *J. Chem. Inf. Model.*, 2019, **59**(8), 3370–3388.
- 23 J. M. Stokes, *et al.*, A Deep Learning Approach to Antibiotic Discovery, *Cell*, 2020, **180**(4), 688–702.
- 24 L. Wu, *et al.*, Relating anatomical therapeutic indications by the ensemble similarity of drug sets, *J. Chem. Inf. Model.*, 2013, **53**(8), 2154–2160.
- 25 K. Gallo, *et al.*, SuperPred 3.0: drug classification and target prediction—a machine learning approach, *Nucleic Acids Res.*, 2022, **50**(W1), W726–W731.
- 26 C. Knox, *et al.*, DrugBank 6.0: the DrugBank Knowledgebase for 2024, *Nucleic Acids Res.*, 2024, **52**(D1), D1265–D1275.
- 27 M. Kuhn, *et al.*, The SIDER database of drugs and side effects, *Nucleic Acids Res.*, 2016, **44**(D1), D1075–D1079.
- 28 A. Rutz, *et al.*, The LOTUS initiative for open knowledge management in natural products research, *eLife*, 2022, **11**, e70780.
- 29 H. Zhao, *et al.*, NPASS database update 2023: quantitative natural product activity and species source database for biomedical research, *Nucleic Acids Res.*, 2023, **51**(D1), D621–D628.
- 30 K. Gao, *et al.*, HERB 2.0: an updated database integrating clinical and experimental evidence for traditional Chinese medicine, *Nucleic Acids Res.*, 2025, **53**(D1), D1404–D1414.
- 31 S. K. Kim, *et al.*, TM-MC 2.0: an enhanced chemical database of medicinal materials in Northeast Asian traditional medicine, *BMC Complementary Med. Ther.*, 2024, **24**(1), 40.
- 32 F.-Y. Sun, *et al.*, InfoGraph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization, *arXiv*, 2019, preprint, arXiv:1908.01000, DOI: [10.48550/arXiv.1908.01000](https://doi.org/10.48550/arXiv.1908.01000).
- 33 Z. Zhu, *et al.*, Torchdrug: A powerful and flexible machine learning platform for drug discovery, *arXiv*, 2022, preprint, arXiv:2202.08320, DOI: [10.48550/arXiv.2202.08320](https://doi.org/10.48550/arXiv.2202.08320).
- 34 J. Nickel, *et al.*, SuperPred: update on drug classification and target prediction, *Nucleic Acids Res.*, 2014, **42**, W26–W31.
- 35 M. Dunkel, *et al.*, SuperPred: drug classification and target prediction, *Nucleic Acids Res.*, 2008, **36**, W55–W59.
- 36 R. A. Hodos, *et al.*, In silico methods for drug repurposing and pharmacology, *Wiley Interdiscip. Rev.: Syst. Biol. Med.*, 2016, **8**(3), 186–210.
- 37 Q. Zheng, *et al.*, Ephedrae herba: A comprehensive review of its traditional uses, phytochemistry, pharmacology, and toxicology, *J. Ethnopharmacol.*, 2023, **307**, 116153.
- 38 Q.-X. Zheng, *et al.*, Review of Rhubarbs: Chemistry and Pharmacology, *Chin. Herb. Med.*, 2013, **5**(1), 9–32.
- 39 Z. M. Li, S. W. Xu and P. Q. Liu, *Salvia miltiorrhiza* Burge (Danshen): a golden herbal medicine in cardiovascular therapeutics, *Acta Pharmacol. Sin.*, 2018, **39**(5), 802–824.



- 40 E. L. Willighagen, *et al.*, The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching, *J. Cheminf.*, 2017, **9**(1), 33.
- 41 S. Byun, *et al.*, Src kinase is a direct target of apigenin against UVB-induced skin inflammation, *Carcinogenesis*, 2013, **34**(2), 397–405.
- 42 X. Yan, *et al.*, Apigenin in cancer therapy: anti-cancer effects and mechanisms of action, *Cell Biosci.*, 2017, **7**, 50.
- 43 T. H. Pham, *et al.*, Apigenin, a Partial Antagonist of the Estrogen Receptor (ER), Inhibits ER-Positive Breast Cancer Cell Proliferation through Akt/FOXO1 Signaling, *Int. J. Mol. Sci.*, 2021, **22**(1), 470.
- 44 X. Long, *et al.*, Apigenin inhibits antiestrogen-resistant breast cancer cell growth through estrogen receptor-alpha-dependent and estrogen receptor-alpha-independent mechanisms, *Mol. Cancer Ther.*, 2008, **7**(7), 2096–2108.
- 45 M. Dean, *et al.*, The Flavonoid Apigenin Is a Progesterone Receptor Modulator with In Vivo Activity in the Uterus, *Horm. Cancer*, 2018, **9**(4), 265–277.
- 46 Z. Javed, *et al.*, Apigenin role as cell-signaling pathways modulator: implications in cancer prevention and treatment, *Cancer Cell Int.*, 2021, **21**(1), 189.
- 47 Y. T. Huang, *et al.*, Inhibitions of protein kinase C and proto-oncogene expressions in NIH 3T3 cells by apigenin, *Eur. J. Cancer*, 1996, **32A**(1), 146–151.
- 48 J. Fu, *et al.*, Apigenin suppresses tumor angiogenesis and growth via inhibiting HIF-1 $\alpha$  expression in non-small cell lung carcinoma, *Chem. Biol. Interact.*, 2022, **361**, 109966.
- 49 S. Freitas, *et al.*, Flavonoids inhibit angiogenic cytokine production by human glioma cells, *Phytother. Res.*, 2010, **25**(6), 916–921.
- 50 W. Zhang, *et al.*, Apigenin inhibits tumor angiogenesis by hindering microvesicle biogenesis via ARHGEF1, *Cancer Lett.*, 2024, **596**, 216961.
- 51 A. J. Kiraly, *et al.*, Apigenin inhibits COX-2, PGE2, and EP1 and also initiates terminal differentiation in the epidermis of tumor bearing mice, *Prostaglandins, Leukotrienes Essent. Fatty Acids*, 2016, **104**, 44–53.
- 52 J. H. Lee, *et al.*, Anti-inflammatory mechanisms of apigenin: inhibition of cyclooxygenase-2 expression, adhesion of monocytes to human umbilical vein endothelial cells, and expression of cellular adhesion molecules, *Arch. Pharmacol Res.*, 2007, **30**(10), 1318–1327.
- 53 Y. C. Liang, *et al.*, Suppression of inducible cyclooxygenase and inducible nitric oxide synthase by apigenin and related flavonoids in mouse macrophages, *Carcinogenesis*, 1999, **20**(10), 1945–1952.
- 54 N. Deng, *et al.*, Anticancer effects of licochalcones: A review of the mechanisms, *Front. Pharmacol.*, 2023, **14**, 1074506.
- 55 S. O. Lee, *et al.*, Licochalcone C Inhibits the Growth of Human Colorectal Cancer HCT116 Cells Resistant to Oxaliplatin, *Biomol. Ther.*, 2024, **32**(1), 104–114.
- 56 Q. Luo, *et al.*, Phillyrin improves myocardial remodeling in salt-sensitive hypertensive mice by reducing endothelin1 signaling, *J. Pharm. Pharmacol.*, 2024, **76**(6), 672–680.
- 57 W. Liu, *et al.*, Phillygenin, a lignan compound, inhibits hypertension by reducing PLC $\beta$ 3-dependent Ca<sup>2+</sup> oscillation, *J. Funct. Foods*, 2019, **60**, 103432.
- 58 Z. Fang, *et al.*, Phillyrin restores metabolic disorders in mice fed with high-fat diet through inhibition of interleukin-6-mediated basal lipolysis, *Front. Nutr.*, 2022, **9**, 956218.
- 59 M. T. Do, *et al.*, Phillyrin attenuates high glucose-induced lipid accumulation in human HepG2 hepatocytes through the activation of LKB1/AMP-activated protein kinase-dependent signalling, *Food Chem.*, 2013, **136**(2), 415–425.
- 60 T. Li, *et al.*, Phillyrin ameliorates DSS-induced colitis in mice via modulating the gut microbiota and inhibiting the NF-kappaB/MLCK pathway, *Microbiol. Spectrum*, 2025, **13**(2), e0200624.
- 61 J. Ma, *et al.*, Phillyrin: A potential therapeutic agent for osteoarthritis via modulation of NF-kappaB and Nrf2 signaling pathways, *Int. Immunopharmacol.*, 2024, **141**, 112960.
- 62 A. M. Diaz Lanza, *et al.*, Lignan and phenylpropanoid glycosides from *Phillyrea latifolia* and their in vitro anti-inflammatory activity, *Planta Med.*, 2001, **67**(3), 219–223.
- 63 S. Zhang, *et al.*, Phillyrin ameliorates influenza a virus-induced pulmonary inflammation by antagonizing CXCR2 and inhibiting NLRP3 inflammasome activation, *Virology*, 2023, **20**(1), 262.

