BMC Bioinformatics

RESEARCH Open Access

Prediction of drug's anatomical therapeutic chemical (ATC) code by constructing biological profiles of ATC codes

Lei Chen[1*], Yiwen Lu[1], Jing Xu[1] and Bo Zhou[2]

*Correspondence:
lchen@shmtu.edu.cn

[1] College of Information Engineering, Shanghai Maritime University, Shanghai 201306, People's Republic of China
[2] School of Basic Medical Sciences, Shanghai University of Medicine and Health Sciences, Shanghai 201318, People's Republic of China

Abstract

Background: The Anatomical Therapeutic Chemical (ATC) classification system, proposed and maintained by the World Health Organization, is among the most widely used drug classification schemes. Recently, it has become a key research focus in drug repositioning. Computational models often pair drugs with ATC codes to explore drug-ATC code associations. However, the limited information available for ATC codes constrains these models, leaving significant room for improvement.

Results: This study presents an inference method to identify highly related target proteins, structural features, and side effects for each ATC code, constructing comprehensive biological profiles. Association networks for target proteins, structural features, and side effects are established, and a random walk with restart algorithm is applied to these networks to extract raw associations. A permutation test is then conducted to exclude false positives, yielding robust biological profiles for ATC codes. These profiles are used to construct new ATC code kernels, which are integrated with ATC code kernels from the existing model PDATC-NCPMKL. The recommendation matrix is subsequently generated using the procedures of PDATC-NCPMKL. Cross-validation results demonstrate that the new model achieves AUROC and AUPR values exceeding 0.96.

Conclusion: The proposed model outperforms PDATC-NCPMKL and other previous models. Analysis of the contributions of the newly added ATC code kernels confirms the value of biological profiles in enhancing the prediction of drug-ATC code associations.

Keywords: Anatomical therapeutic chemical code, Drug repositioning, Biological profiles, Random walk with restart, Network consistency projection

Introduction

Drug development activities are characterized by high risk, substantial financial investment, and an extended research and development (R&D) cycle [1]. Despite significant investments, the success rate of drug development programs remains low [2]. Recently, drug repositioning has emerged as a prominent area of focus due to its potential to accelerate the process of discovering "new drugs," which highlights the novel effects of existing medications [3]. Existing drugs have undergone extensive

clinical toxicological testing, resulting in relatively lower investment and risk when compared to newly designed drugs. However, identifying the new effects of existing drugs is challenging. Traditional methods primarily rely on the expertise of healthcare professionals, which is often insufficient given the vast array of available drugs. Thus, designing effective computational methods has become a promising and popular alternative in recent years [1, 3, 4].

The Anatomical Therapeutic Chemical (ATC) classification system, developed and maintained by the World Health Organization (WHO), serves as a globally standardized drug classification framework that is crucial for international drug use research [5]. This system comprises five levels, with the first level containing fourteen distinct classes, whereas additional classes are identified in subsequent levels. A detailed stratification of ATC codes is available in Additional file 1. Each drug within this system is assigned at least one ATC code, which consists of letters or numbers corresponding to the first through fifth levels. These designations indicate the essential properties of the drugs at various levels. When the ATC codes of a drug or chemical are determined across different levels, they provide insights into its potential therapeutic and pharmacological effects, which can be substantiated through rigorous experimentation. Consequently, research related to this classification system has emerged as a significant area of interest in drug repositioning. As highlighted, computational methods have gained popularity in drug repositioning in recent years. Several computational approaches have been proposed for the ATC classification system. Based on the classification models established in prior studies, these computational methods can be broadly categorized into two groups. The first group focuses solely on the first level of ATC codes, which includes fourteen letters. In this context, the letters represent classes, and the drugs are treated as samples. Since multiple drugs can belong to more than one class, studies within this category have designed multi-label classifiers to categorize drugs into these fourteen classes [6–22]. Accurate representations of drugs are critical for the effectiveness of these classifiers. Early classifiers primarily relied on structural similarities, drug interactions, and drug ontology to develop their classification models. Subsequently, advanced feature extraction and enhancement techniques—including network embedding algorithms, graph transformer networks, graph convolutional networks, and convolutional neural networks—were employed to generate more informative drug features. However, a notable limitation of the aforementioned studies is their inability to fully identify the complete ATC codes for the drugs in question. To address this limitation, some studies have treated drug-ATC code pairs as samples, effectively converting the task of predicting drug ATC codes into the identification of drug-ATC code associations [23–27]. These studies constitute the second category of research. The primary challenge faced by these studies lies in the need to thoroughly and accurately evaluate ATC codes. The previous studies have utilized the letters and numbers within ATC codes, the drugs associated with each ATC code, or the hierarchical structure of the ATC code tree to assess the associations between ATC codes. However, this limited information does not fully reflect the essential characteristics of the ATC codes, leaving significant room for improvement. Furthermore, other studies have approached the drug ATC classification system differently, employing network propagation methods to uncover the associations

between drugs and ATC codes [28–30]. Nonetheless, the learning capabilities of these methods are somewhat restricted, as they lack a systematic training process.

This study focused on identifying associations between drugs and ATC codes. As previously mentioned, existing research struggles to accurately assess these associations due to limited information on ATC codes. To address this limitation, an inference method was developed in this study to identify related target proteins, structural features, and side effects for each ATC code. This method employed the Random Walk with Restart (RWR) algorithm [31, 32] and applied it to networks that included drugs, ATC codes, and the associated target proteins/structural features/side effects. A permutation test was performed subsequently to control for false positives. The resulting target proteins, structural features, and side effects of the ATC codes, termed biological profiles, were used to construct new ATC code kernels. These new kernels were integrated with the existing ATC code kernels from our previous model, PDATC-NCPMKL [27], to create a unified ATC code kernel. Following the same procedures as in PDATC-NCPMKL, an updated model named PDATC-NCPMKL-updated was established. The cross-validation results demonstrated the model's high performance, outperforming PDATC-NCPMKL and other preceding models. The contribution of the newly constructed ATC code kernels to the model's development was analyzed, confirming the beneficial role of the biological profiles of ATC codes in predicting drug-ATC code associations.

Materials and methods

Data on drug-ATC code association

This study utilized drug-ATC code associations previously collected in a past study [27]. Specifically, these associations were retrieved from DrugBank (https://go.drugbank.com/) [33, 34], a public database that compiles comprehensive and reliable drug data. A total of 2930 drugs (with ECFP fingerprint) and their ATC codes at the second, third, and fourth levels of the drug ATC classification system were included. Corresponding to the ATC codes of these drugs, a total of 2930 drugs were paired with ATC codes across three levels, resulting in three distinct drug-ATC code association groups. In detail, for the ATC codes at the second level, there were 3619 drug-ATC code associations covering 86 ATC codes. For the ATC codes at the third level, 3886 drug-ATC code associations were established for 231 ATC codes. Regarding the fourth level, 720 ATC codes were involved, which accounted for 4133 drug-ATC code associations. For each of the three levels, a dataset was constructed using the aforementioned drug-ATC code associations as positive samples, whereas randomly paired drugs and ATC codes were considered negative samples, with equal numbers of positive and negative samples. The three datasets were denoted as DA_2, DA_3, and DA_4, where the subscript indicates the levels of ATC codes. Furthermore, the drug-ATC code associations at the (i)-th level can be represented by an association adjacency matrix, denoted as F^i. The value of $F^i(j,k)$ is set to 1 if and only if the (j)-th drug and the (k)-th ATC code form an association.

This study initially inferred the biological profiles of ATC codes and utilized them to construct a model for predicting drug-ATC code associations. In building reliable biological profiles of ATC codes, additional drug-ATC code associations were included, including drugs that do not have ECFP fingerprint. This approach enhances the reliability of the biological profiles of ATC codes. These associations were obtained from

Table 1 Details of drug-ATC code associations at the second, third and fourth levels for inferring biological profiles of ATC codes

Level	Number of ATC codes	Number of drugs	Number of associations
Second	90	3396	4128
Third	245	3396	4400
Fourth	784	3396	4666

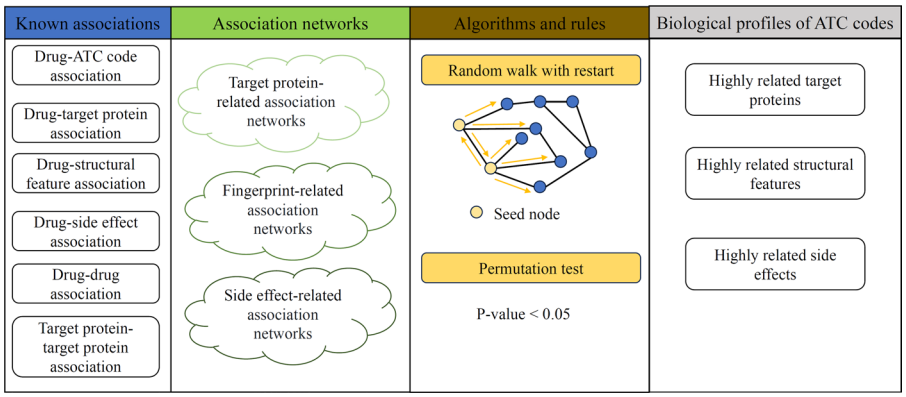

Fig. 1 Procedures for inferring biological profiles of ATC codes. Six different association types are employed to construct target protein-related, fingerprint-related, and side effect-related association networks. The random walk with restart algorithm is adopted to infer related target proteins, structural features, and side effects for each ATC code, followed by a permutation test to control false positives. The final biological profiles of ATC codes include highly related target proteins, structural features, and side effects

DrugBank. Table 1 provides the counts of drugs, ATC codes, and drug-ATC code associations at the second, third, and fourth levels, specifically 4128, 4400, and 4666 associations, respectively. This data covers 90, 245, and 784 ATC codes across 3,396 drugs. Among these 3,396 drugs, 2,934 were small molecules (2930 drugs with ECFP fingerprints, remaining four drugs without ECFP fingerprints), remaining 462 were protein drugs. Above three drug-ATC code association groups constituted three datasets, denoted by DA_2^T, DA_3^T, and DA_4^T.

Inference of ATC codes' biological profiles

In this study, the prediction of ATC codes for drugs was based on identifying drug-ATC code associations. ATC codes represent the core components of all samples, making it crucial to integrate their essential information when constructing the prediction model. However, the existing information on ATC codes is quite limited. To address this, a computational method was developed to infer the biological profiles of ATC codes, which would be utilized in constructing the model for predicting drug-ATC code associations. Specifically, associations between ATC codes and target proteins, structural features, or drug side effects were inferred. The entire procedure is illustrated in Fig. 1.

Association data for inferring ATC codes' biological profiles

In addition to the drug-ATC code associations described in Section "Data on drug-ATC code association" (datasets DA_2^T, DA_3^T, and DA_4^T), additional data on

drug-target protein associations, drug-structural feature associations, drug-side effect associations, drug-drug associations, and target protein-target protein associations were also incorporated, as detailed below.

Drug-target protein association The target proteins for all drugs were sourced from DrugBank [33, 34]. After limiting the focus to 3396 drugs, a total of 10,091 drug-target protein associations were identified. These associations included 2048 drugs and 2435 target proteins.

Drug-structural feature association Drugs can typically be represented using the Simplified Molecular Input Line Entry System (SMILES) format [35]. In this study, the SMILES strings for 2934 of the 3396 drugs were collected from DrugBank and subsequently input into RDKit (http://www.rdkit.org/), an open-source cheminformatics and machine learning toolkit, to derive their ECFP fingerprints [36]. This process yielded 122,374 drug-structural feature associations, covering 1024 structural features. Above associations involved 2930 drugs as not all 2934 drugs have structural features.

Drug-side effect association This type of association was sourced from SIDER (http://sideeffects.embl.de, version 4.1) [37]. The file "meddra_all_se.tsv" was downloaded, containing the side effects for all listed drugs. After narrowing the selection to the intersection of 3396 drugs and 1430 drugs in SIDER, 937 drugs were obtained. A total of 113,474 drug-side effect associations for 937 drugs were identified, including 5464 distinct side effects.

Drug-drug association STITCH (http://stitch4.embl.de/) [38, 39] is a public database that compiles various information regarding chemical substances, including chemical-chemical interactions, chemical-protein interactions, and chemical SMILES strings. Drug-drug associations were extracted from the file "chemical_chemical.links.detailed.v4.0.tsv" within this database. This file contains numerous chemical-chemical interactions, represented by PubChem IDs. Furthermore, each interaction is quantified by a confidence score, referred to as "Combined_score", which ranges from 1 to 999. The chemical-chemical interactions for the 3396 drugs led to the extraction of 133,772 drug-drug associations, and the corresponding confidence scores for these associations were also obtained. For drugs d_1 and d_2, the confidence score of the association between them is denoted as $CS_d(d_1, d_2)$.

Target protein-target protein association STRING (https://cn.string-db.org/) [40] is a public database that gathers information on known and predicted protein–protein interactions. These interactions illustrate both direct (physical) and indirect (functional) associations between proteins. The file "9606.protein.links.detailed.v12.0.txt" includes a vast array of protein–protein interactions, with proteins represented by Ensembl IDs. Like chemical-chemical interactions, each protein–protein interaction is assigned a confidence score, also termed "Combined_score", ranging from 1 to 999. The extracted protein–protein interactions involved 2435 target proteins, resulting in 1746 target protein-target protein associations. The confidence scores for these associations were utilized in this study. To denote the confidence score of the association between proteins p_1 and p_2, we use the notation $CS_p(p_1, p_2)$.

The details of the aforementioned five association types are summarized in Table 2.

Table 2 Details of five association types

Association type	Number of objects	Number of associations
Drug-target protein	2048 drugs, 2435 target proteins	10,091
Drug-structural feature	2930 drugs, 1024 structural features	122,374
Drug-side effect	937 drugs, 5464 side effects	113,474
Drug-drug	3396 drugs	133,772
Target protein-target protein	2435 target proteins	1746

Association network construction

Based on the aforementioned associations, several association networks were constructed to infer the target proteins, structural features, and side effects associated with each ATC code. The following notations are introduced: Let n represent the number of drugs, which together form a drug set $D = \{d_1, d_2, \ldots, d_n\}$. The number of ATC codes at the i-th level is denoted by m_i, with the set $A_i = \{a_1^i, a_2^i, \ldots, a_{m_i}^i\}$ containing these ATC codes. Let p be the number of target proteins within the target protein set $T = \{t_1, t_2, \ldots, t_p\}$. The number of structural features is indicated by r, and the corresponding set of these structural features is represented by $F = \{f_1, f_2, \ldots, f_r\}$. Let q denote the number of side effects, with the set $S = \{s_1, s_2, \ldots, s_q\}$ including these side effects. Three groups of association networks were constructed for target proteins, structural features, and side effects, each group comprising three association networks corresponding to the three ATC code levels. The construction procedures are outlined as follows.

Target protein-related association networks For ATC codes at the second, third, and fourth levels, a target protein-related association network was constructed for each level. Each network was established in a similar manner, defining drugs, target proteins, and ATC codes at the respective level as nodes—collectively known as the node set $V_{TP}^i = D \cup T \cup A_i$. The edges within this network were determined based on drug-ATC code, drug-target protein, drug-drug, and target protein-target protein associations. Additionally, each edge was assigned a weight. For an edge corresponding to a drug-drug association, the weight was defined as $\frac{CS_d(d_1,d_2)}{1000}$, where $CS_d(d_1, d_2)$ is the confidence score of the drug-drug association. For edges corresponding to target protein-target protein associations, the weight was defined as $\frac{CS_p(p_1,p_2)}{1000}$, where $CS_p(p_1, p_2)$ is the confidence score for that association. The weights for the remaining edges were set to one. For ease of reference, the association networks in this group are denoted by N_{TP}^i, where $i \in \{2, 3, 4\}$ represents the ATC code level. Detailed information regarding three target protein-related networks is provided in Table 3. A brief illustration of one target protein-related network is displayed in Fig. 2A.

Fingerprint-related association networks Likewise, three fingerprint-related association networks were constructed, corresponding to three ATC code levels, using a similar setup. These networks included drugs, structural features from the ECFP fingerprint, and ATC codes at the second, third, or fourth levels as nodes, forming the node set $V_{FP}^i = D \cup F \cup A_i$. The edges represented drug-ATC code, drug-structural feature, and drug-drug associations, and each was assigned a weight. The weight of an edge

Table 3 Details of target protein-related association networks

ATC code level	Number of nodes				Number of edges				
	Drugs	ATC codes	Target proteins	Total	Drug-Drug	Drug-ATC code	Drug-target protein	Target protein-target protein	Total
Second	3396	90	2435	5921	133,772	4128	10,091	1746	149,737
Third	3396	245	2435	6076	133,772	4400	10,091	1746	150,009
Fourth	3396	784	2435	6615	133,772	4666	10,091	1746	150,275

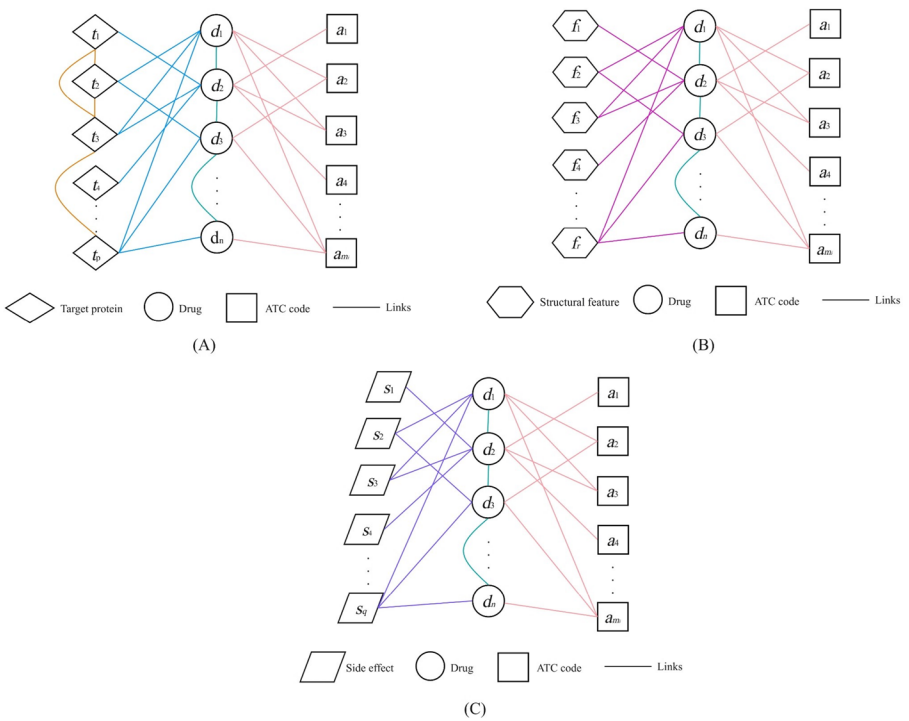

Fig. 2 An illustration on the network in each group. **A** Target protein-related association network; **B** Fingerprint-related association network; **C** Side effect-related association network

representing a drug-drug association was defined as $\frac{CS_d(d_1,d_2)}{1000}$, where $CS_d(d_1, d_2)$ denotes the confidence score of the association. For other edges, weights were set to one. These networks were denoted by N^i_{FP}, where $i \in \{2, 3, 4\}$ corresponds to the ATC code level. Detailed information on the three fingerprint-related networks is provided in Table 4, whereas a brief illustration of one target protein-related network is shown in Fig. 2B.

Side effect-related association networks The final network group comprised three networks corresponding to three ATC code levels. These networks included drugs, ATC codes at the second, third, or fourth levels, and side effects as nodes, forming the node set $V^i_{SE} = D \cup S \cup A_i$. Edges represented drug-ATC code, drug-drug, and drug-side effect associations, and each was assigned a weight. For edges representing drug-drug associations, weights were defined as in the previously described networks. All other edges were assigned a weight of one. These networks were denoted by N^i_{SE}, where

Table 4 Details of fingerprint-related association networks

ATC code level	Number of nodes				Number of edges			
	Drugs	ATC codes	Structural features	Total	Drug-Drug	Drug-ATC code	Drug-structural feature	Total
Second	3396	90	1024	4510	133,772	4128	122,374	260,274
Third	3396	245	1024	4665	133,772	4400	122,374	260,546
Fourth	3396	784	1024	5204	133,772	4666	122,374	260,812

Table 5 Details of side effect-related association networks

ATC code level	Number of nodes				Number of edges			
	Drugs	ATC codes	Side effects	Total	Drug-Drug	Drug-ATC code	Drug-side effect	Total
Second	3396	90	5464	8950	133,772	4128	113,474	251,374
Third	3396	245	5464	9105	133,772	4400	113,474	251,646
Fourth	3396	784	5464	9644	133,772	4666	113,474	251,912

$i \in \{2, 3, 4\}$ indicates the ATC code level. Detailed information on three side effect-related networks is provided in Table 5, and a brief illustration of one target protein-related network is shown in Fig. 2C.

Random walk with restart algorithm

The constructed networks connect ATC codes to three elements (target proteins, structural features, and side effects), utilizing drugs as intermediaries. The latent associations between ATC codes and these three elements can be extracted using appropriate network algorithms. In this study, the RWR algorithm [31, 32] was employed to extract such associations. The RWR algorithm has extensive applications in identifying disease-related genes [41–44]. Given a network N, the RWR algorithm simulates a random walker starting from one or more nodes and moves through the network to propagate probabilities from seed nodes to other nodes. If there are m seed nodes, the initial probability assigned to each seed node is defined as $1/m$. The probabilities of all other nodes are initially set to zero. These probabilities are represented in a probability vector, denoted as P_0. The RWR algorithm iteratively updates the probability vector as follows:

$$P_{t+1} = (1 - r)A^T P_t + rP_0, \tag{1}$$

where A is the column-wise normalized adjacency matrix of network N and r is the restarting probability, which was set to 0.7 in this study. If the probability vectors P_t and P_{t+1} are close enough, which is measured by $\|P_{t+1} - P_t\|_{L_1} < 10^{-10}$, the updating procedure stops. P_{t+1} is picked up as the outcome of RWR algorithm. Based on this probability vector, each node, excluding the seed nodes, is assigned a probability that reflects the association between that node and the seed nodes. A higher probability indicates a stronger association.

In this study, the RWR algorithm was implemented for each network as detailed in Section "Association network construction". For each ATC code, the node corresponding to that ATC code, along with the drug nodes labeled by it, were selected as seed nodes. Utilizing these seed nodes within the RWR algorithm assists in identifying relevant target proteins, structural features, and side effects associated with the corresponding ATC code.

Permutation test

Although the RWR algorithm is effective in uncovering hidden linkages within the network, its accuracy is not guaranteed. The resulting probability vector can be influenced by the network's structure. Certain nodes within the network, such as cut vertices, are more likely to receive high probabilities, regardless of the selected seed nodes, due to the necessity of multiple paths passing through them. However, these nodes may not always have significant relevance to the seed nodes. To mitigate the impact of this factor, a permutation test was devised.

For ATC code a_l^i, where i represented the ATC code level, k drugs were labeled by this ATC code. The RWR algorithm with a_l^i and these k drugs as seed nodes was executed on the networks N_{TP}^i, N_{FP}^i, and N_{SE}^i. Each target protein t_j, structural feature f_j and side effect s_j was assigned a probability, denoted by $P(t_j)$, $P(f_j)$, and $P(s_j)$, respectively. The permutation test first randomly constructed 500 node sets, each of which contained one ATC code node and same number of drug nodes, which were denoted by $D_1, D_2, \ldots, D_{500}$. Then, nodes in each randomly produced node set were fed into RWR algorithm as seed nodes. After that, the target protein t_j, structural feature f_j and side effect s_j were also assigned a probability under each randomly produced node set. They were denoted by $P_{D_k}(t_j)$, $P_{D_k}(f_j)$, and $P_{D_k}(s_j)$, where $1 \leq k \leq 500$. The significance of $P(t_j)$ was evaluated by comparing it and $P_{D_k}(t_j)$ $(k = 1, 2, \ldots, 500)$, which was measured by

$$P - \text{value}(t_j) = \frac{\left|\{D_k : P_{D_k}(t_j) > P(t_j),\ 1 \leq k \leq 500\}\right|}{500}, \tag{2}$$

If $P - \text{value}(t_j)$ was small, $P(t_j)$ was significantly larger than the probabilities on randomly produced node sets. In this case, target protein t_j was deemed highly related to ATC code a_l^i. Here, 0.05 was selected as the threshold as it is always the cutoff of statistical significance. Accordingly, the highly related target proteins of ATC code a_l^i can be obtained, which constituted the set $TP(a_l^i)$. The highly related structural features and side effects of ATC code a_l^i can be accessed in the same manner. They constituted the highly related structural feature set $FP(a_l^i)$ and highly related side effect set $SE(a_l^i)$.

ATC code kernel construction

In Section "Inference of ATC codes' biological profiles", a computational method was developed to infer the highly related target proteins, structural features, and side effects for each ATC code. Based on these findings, the ATC code kernels were constructed. For ATC code a_l^i, the one-hot scheme was applied to $TP(a_l^i)$, $FP(a_l^i)$, and $SE(a_l^i)$ for generating the feature representations of a_l^i, formulated by $V_{TP}(a_l^i)$, $V_{FP}(a_l^i)$, and $V_{SE}(a_l^i)$. For three above ATC code representations, the following five kernel functions were employed to construct the kernels of ATC codes.

$$\begin{cases} K^i_{GIP-TP,a}\left(a^i_j, a^i_k\right) = \exp\left(-\gamma\left\|V_{TP}\left(a^i_j\right) - V_{TP}\left(a^i_k\right)\right\|^2\right) \\ K^i_{Corr-TP,a}\left(a^i_j, a^i_k\right) = \dfrac{Cov\left(V_{TP}\left(a^i_j\right), V_{TP}\left(a^i_k\right)\right)}{\sqrt{Var\left(V_{TP}\left(a^i_j\right)\right)Var\left(V_{TP}\left(a^i_k\right)\right)}} \\ K^i_{COS-TP,a}\left(a^i_j, a^i_k\right) = \dfrac{V_{TP}\left(a^i_j\right)V_{TP}\left(a^i_k\right)^T}{\left|V_{TP}\left(a^i_j\right)\right|\left|V_{TP}\left(a^i_k\right)\right|} \\ K^i_{MI-TP,a}\left(a^i_j, a^i_k\right) = \sum_{u=0}^{1}\sum_{v=0}^{1} f(u,v)\log\dfrac{f(u,v)}{f(u)f(v)} \\ K^i_{Jaccard-TP,a}\left(a^i_j, a^i_k\right) = \dfrac{\left|N_{a^i_j} \cap N_{a^i_k}\right|}{\left|N_{a^i_j} \cup N_{a^i_k}\right|} \end{cases}, \tag{3}$$

$$\begin{cases} K^i_{GIP-FP,a}\left(a^i_j, a^i_k\right) = \exp\left(-\gamma\left\|V_{FP}\left(a^i_j\right) - V_{FP}\left(a^i_k\right)\right\|^2\right) \\ K^i_{Corr-FP,a}\left(a^i_j, a^i_k\right) = \dfrac{Cov\left(V_{FP}\left(a^i_j\right), V_{FP}\left(a^i_k\right)\right)}{\sqrt{Var\left(V_{FP}\left(a^i_j\right)\right)Var\left(V_{FP}\left(a^i_k\right)\right)}} \\ K^i_{COS-FP,a}\left(a^i_j, a^i_k\right) = \dfrac{V_{FP}\left(a^i_j\right)V_{FP}\left(a^i_k\right)^T}{\left|V_{FP}\left(a^i_j\right)\right|\left|V_{FP}\left(a^i_k\right)\right|} \\ K^i_{MI-FP,a}\left(a^i_j, a^i_k\right) = \sum_{u=0}^{1}\sum_{v=0}^{1} f(u,v)\log\dfrac{f(u,v)}{f(u)f(v)} \\ K^i_{Jaccard-FP,a}\left(a^i_j, a^i_k\right) = \dfrac{\left|N_{a^i_j} \cap N_{a^i_k}\right|}{\left|N_{a^i_j} \cup N_{a^i_k}\right|} \end{cases}, \tag{4}$$

$$\begin{cases} K^i_{GIP-SE,a}\left(a^i_j, a^i_k\right) = \exp\left(-\gamma\left\|V_{SE}\left(a^i_j\right) - V_{SE}\left(a^i_k\right)\right\|^2\right) \\ K^i_{Corr-SE,a}\left(a^i_j, a^i_k\right) = \dfrac{Cov\left(V_{SE}\left(a^i_j\right), V_{SE}\left(a^i_k\right)\right)}{\sqrt{Var\left(V_{SE}\left(a^i_j\right)\right)Var\left(V_{SE}\left(a^i_k\right)\right)}} \\ K^i_{COS-SE,a}\left(a^i_j, a^i_k\right) = \dfrac{V_{SE}\left(a^i_j\right)V_{SE}\left(a^i_k\right)^T}{\left|V_{SE}\left(a^i_j\right)\right|\left|V_{SE}\left(a^i_k\right)\right|} \\ K^i_{MI-SE,a}\left(a^i_j, a^i_k\right) = \sum_{u=0}^{1}\sum_{v=0}^{1} f(u,v)\log\dfrac{f(u,v)}{f(u)f(v)} \\ K^i_{Jaccard-SE,a}\left(a^i_j, a^i_k\right) = \dfrac{\left|N_{a^i_j} \cap N_{a^i_k}\right|}{\left|N_{a^i_j} \cup N_{a^i_k}\right|} \end{cases}, \tag{5}$$

where γ was the Gaussian kernel bandwidth (it was set to 1 in this study), $u, v \in \{0, 1\}$, $f(u)/f(v)$ was the observed frequency of value u in $V_{TP}\left(a_j^i\right)/V_{TP}\left(a_k^i\right)$, $V_{FP}\left(a_j^i\right)/V_{FP}\left(a_k^i\right)$, or $V_{SE}\left(a_j^i\right)/V_{SE}\left(a_k^i\right)$, $f(u,v)$ denoted the observed relative frequency of (u,v) in $V_{TP}\left(a_j^i\right)$ and $V_{TP}\left(a_k^i\right)$, $V_{FP}\left(a_j^i\right)$ and $V_{FP}\left(a_k^i\right)$, or $V_{SE}\left(a_j^i\right)$ and $V_{SE}\left(a_k^i\right)$, $N_{a_j^i}/N_{a_k^i}$ was a set containing the components with value 1 in $V_{TP}\left(a_j^i\right)/V_{TP}\left(a_k^i\right)$, $V_{FP}\left(a_j^i\right)/V_{FP}\left(a_k^i\right)$, or $V_{SE}\left(a_j^i\right)/V_{SE}\left(a_k^i\right)$. Accordingly, five kernels were obtained for each representation of ATC codes and totally 15 kernels were accessed, which are listed in Table 6.

The five ATC code kernels derived from the same feature representation were fused by the multiple kernel learning algorithm provided in [27]. Three ATC code kernels were generated, denoted by $K_{TP,a}^{i,*}$, $K_{FP,a}^{i,*}$, and $K_{SE,a}^{i,*}$.

PDATC-NCPMKL-updated for the prediction of drug's ATC code

In reference [27], Chen et al. proposed a recommendation model for predicting drug ATC codes, which constructed various drug and ATC code kernels. However, the existing ATC code kernels may not fully capture the associations due to limited information available for ATC codes. To address this limitation, a computational method was designed to infer related target proteins, structural features, and side effects of ATC codes, thereby providing additional data for the construction of ATC code kernels, as discussed in Section "ATC code kernel construction". Consequently, the PDATC-NCPMKL model was updated by integrating the newly constructed ATC code kernels. The procedures of this updated model are illustrated in Fig. 3, and it is referred to as PDATC-NCPMKL-updated for clarity.

Five drug kernels ($K_{TP,d}^{i,*}$, $K_{FP,d}^{i,*}$, $K_{IN,d}^{i}$, $K_{F,d}^{i,*}$, and $K_{SE,d}^{i,*}$) were constructed in PDATC-NCPMKL, which were derived from drug target proteins, structural features (ECFP fingerprint), interactions, ATC codes, side effects. Their brief introduction is provided in Additional file 2. They were also retained in our model and fused in the same way in PDATC-NCPMKL. The integrated drug kernel was denoted by K_d^i, where i was the ATC code level. On the other hand, three ATC code kernels were built in PDATC-NCPMKL. They were $K_{F,a}^{i,*}$, K_{SPro}^{i}, and K_{SM}^{i}, which were derived from drug-ATC code associations, ATC code tree (see Eq. (2) in Wang et al.'s study [24]), and numbers or letters in ATC codes (see Eq. (2) in Zhao et al.'s study [26]), respectively. Their brief introduction is also available in Additional file 2. In this study, three additional ATC code kernels were constructed: $K_{TP,a}^{i,*}$, $K_{FP,a}^{i,*}$, and $K_{SE,a}^{i,*}$. They were fused with the ATC code kernels in PDATC-NCPMKL as follows:

$$K_a^i = \frac{K_{F,a}^{i,*}+K_{SPro}^{i}+K_{SM}^{i}+K_{TP,a}^{i,*}+K_{FP,a}^{i,*}+K_{SE,a}^{i,*}}{6}, \quad (6).$$

Table 6 Details of ATC code kernels at the i-th level

Kernel function	Target protein	Fingerprint	Side effect
GIP	$K_{GIP-TP,a}^{i}$	$K_{GIP-FP,a}^{i}$	$K_{GIP-SE,a}^{i}$
Corr	$K_{Corr-TP,a}^{i}$	$K_{Corr-FP,a}^{i}$	$K_{Corr-SE,a}^{i}$
COS	$K_{COS-TP,a}^{i}$	$K_{COS-FP,a}^{i}$	$K_{COS-SE,a}^{i}$
MI	$K_{MI-TP,a}^{i}$	$K_{MI-FP,a}^{i}$	$K_{MI-SE,a}^{i}$
Jaccardscore	$K_{Jaccard-TP,a}^{i}$	$K_{Jaccard-FP,a}^{i}$	$K_{Jaccard-SE,a}^{i}$

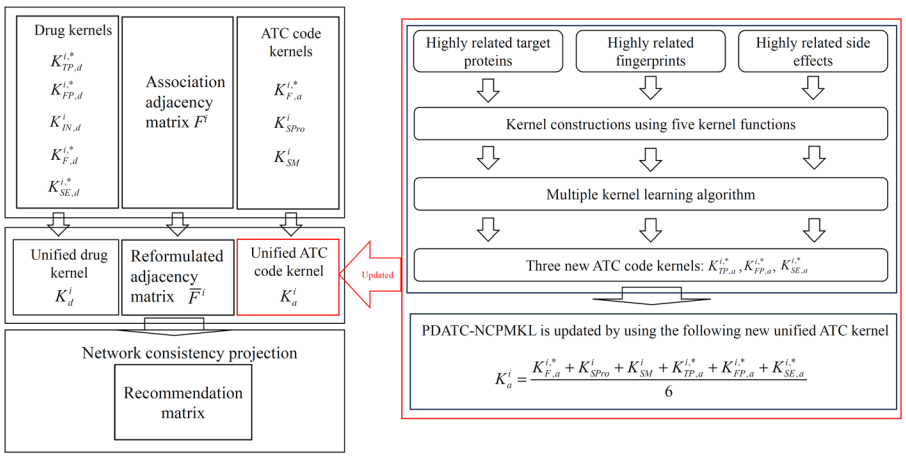

Fig. 3 Procedures of the PDATC-NCPMKL-updated. This model is an updated version of PDATC-NCPMKL. Based on the biological profiles of ATC codes, include highly related target proteins, structural features, side effects, three new ATC code kernels ($K_{TP,a}^{i,*}$, $K_{FP,a}^{i,*}$, and $K_{SE,a}^{i,*}$) are built. These new ATC code kernels are fused with those in PDATC-NCPMKL to yield novel unified ATC code kernel. The recommendation matrix is produced with the same following procedures of PDATC-NCPMKL. Please refer to Additional file 2 for drug kernels ($K_{TP,d}^{i,*}$, $K_{FP,d}^{i,*}$, $K_{IN,d}^{i}$, $K_{F,d}^{i,*}$, and $K_{SE,d}^{i,*}$) and ATC code kernels ($K_{F,a}^{i,*}$, K_{SPro}^{i}, and K_{SM}^{i}) in PDATC-NCPMKL

After that, the association adjacency matrix F^i was processed by the Weighted K Nearest Known Neighbors (WKNKN) [45], producing a generalized adjacency matrix $\overline{F}^i$. Finally, the network consistency projection was applied to $\overline{F}^i$, K_d^i, and K_a^i for generating the recommendation matrix.

Performance evaluation

Cross-validation is a widely recognized method for assessing the performance of predictive models [46]. In this methodology, samples (specifically, drug-ATC code pairs in this study) are randomly and equally divided into several subsets. Each subset is subsequently designated as the test set whereas the remaining subsets form the training set. The model built on the training set is employed to predict the samples in the test set, and the average performance across all test sets is typically calculated to evaluate the model. Generally, samples are divided into five or ten subsets. This study performed ten-fold cross-validation to evaluate the performance of the models.

The results of the cross-validation were quantified using AUROC and AUPR to assess model performance [47–50]. AUROC represents the area under the ROC curve. To generate a ROC curve, various thresholds for identifying positive samples are initially selected. For each threshold, the cross-validation results are classified as true positive (TP), false negative (FN), false positive (FP), and true negative (TN). The true positive rate and false positive rate are calculated as follows:

$$\begin{cases} True\ positive\ rate = \dfrac{TP}{TP + FN} \\ False\ positive\ rate = \dfrac{FP}{FP + TN} \end{cases}. \quad (7).$$

Once a set of true positive rates and false positive rates across different thresholds is established, the ROC curve is plotted with the true positive rate on the Y-axis and

the false positive rate on the X-axis. Similarly, AUPR represents the area under the Precision-Recall (PR) curve and is also calculated using several thresholds for determining positive samples. In this context, recall (equivalent to the true positive rate) and precision at a specific threshold are computed. The formula for calculating precision is expressed as:

$Precision = \frac{TP}{TP+FP}$. (8).

Likewise, after obtaining a set of recall and precision values at varying thresholds, the PR curve is plotted with precision on the Y-axis and recall on the X-axis. The AUROC and AUPR values range between 0 and 1. A higher value indicates better model performance.

Results and discussion

Biological profiles of ATC codes

To address the challenge of limited information regarding ATC codes, a computational method was developed to identify closely related target proteins, structural features, and side effects associated with each ATC code. This method sequentially employed the RWR algorithm and permutation tests. Based on the identified highly related target proteins, structural features, and side effects of the ATC codes, associations were constructed for ATC code-target protein, ATC code-structural feature, and ATC code-side effect, as detailed in Additional files 3, 4, 5. Table 7 provides statistics on these associations, revealing 9987, 15,871, and 32,628 associations between ATC codes at the second, third, and fourth levels, respectively, and their respective target proteins. It is noteworthy that not all ATC codes were associated with highly related target proteins. Figure 4 further depicts the distribution of ATC codes across the three levels based on their corresponding target proteins. Most ATC codes were associated with fewer than 100 target proteins, and the quantity of ATC codes exhibited a decreasing trend as the number of highly related target proteins increased. This finding is logical, reinforcing the reliability of the results. Similar observations regarding ATC code-structural feature and ATC code-side effect associations can be drawn from Table 7 and Fig. 4. Overall, the biological profiles of ATC codes presented in this study appear to be robust and may serve as valuable resources for investigating drug-ATC code associations and addressing other drug-related inquiries.

Table 7 Statistics of biological profiles of ATC codes

Object	ATC code level	Number of ATC codes	Number of objects	Number of associations
Target protein	Second	88	2340	9987
	Third	229	2414	15,871
	Fourth	726	2431	32,628
Fingerprint	Second	85	1020	8703
	Third	230	1024	14,661
	Fourth	715	1024	29,414
Side effect	Second	85	5186	14,751
	Third	224	5390	29,575
	Fourth	649	5463	75,525

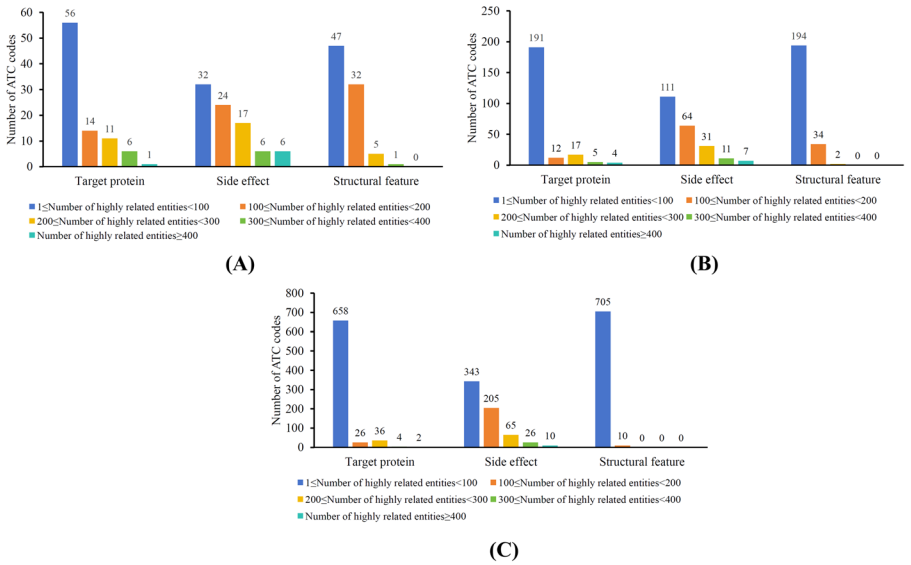

Fig. 4 Bar chart to show the distribution of ATC codes at three levels based on numbers of their highly related target proteins, structural features, and side effects. **A** Bar chart for ATC codes at the second level; **B** Bar chart for ATC codes at the third level; **C** Bar chart for ATC codes at the fourth level. The number above each bar represents the number of ATC codes having corresponding number of highly related entities. For example, 56 in (**A**) suggests that there are 56 ATC codes at the second level having 1–100 highly related target proteins

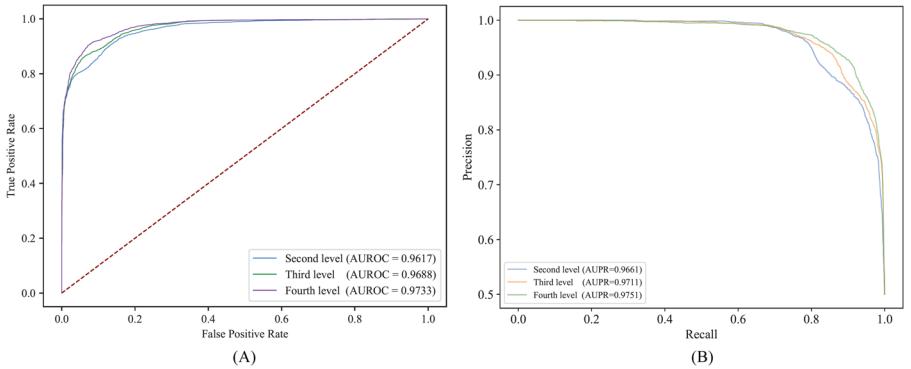

Fig. 5 ROC and PR curves to show the performance of PDATC-NCPMKL-updated. **A** ROC curves; **B** PR curves. The AUROC and AUPR values are all higher than 0.96, suggesting the high performance of PDATC-NCPMKL-updated

Performance of PDATC-NCPMKL-updated

In this study, the previous model PDATC-NCPMKL was enhanced by incorporating newly discovered biological profiles of ATC codes. For consistency, the parameters in PDATC-NCPMKL were retained. Specifically, the weight parameter w in WKNKN, which influences the weights of different neighbors, was set to 0.9. The updated model, PDATC-NCPMKL-updated, was evaluated through ten-fold cross-validation. The ROC and PR curves are presented in Fig. 5, alongside the AUROC and AUPR values. The AUROC values for the second, third, and fourth levels were found to be 0.9617, 0.9688, and 0.9733, respectively. Meanwhile, the AUPR values for the three ATC code levels were 0.9661, 0.9711, and 0.9751. All values exceeded 0.96, indicating the high performance of

the PDATC-NCPMKL-updated model. Furthermore, the model's performance progressively improved with increasing ATC code levels.

Ablation tests on ATC code kernels

The model PDATC-NCPMKL-updated was developed by incorporating new ATC code kernels derived from the recently established biological profiles of ATC codes. This model comprises six ATC code kernels, necessitating an analysis of their individual contributions. To achieve this, each ATC code kernel was sequentially removed, resulting in the creation of six distinct models—each constructed by omitting a single ATC code kernel. These models underwent evaluation using ten-fold cross-validation, and the resulting AUROC and AUPR values are presented in Table 8.

Firstly, when comparing the performance of PDATC-NCPMKL-updated as discussed in Section "Performance of PDATC-NCPMKL-updated", it is evident that each of the aforementioned models produced lower AUROC and AUPR values, indicating that they were less effective than PDATC-NCPMKL-updated. Since these models were devised by removing one ATC code kernel from the original model, each ATC code kernel contributed positively to the performance of PDATC-NCPMKL-updated, including the new ATC code kernels developed in this research. This further demonstrates that the newly created biological profiles of ATC codes are valuable for predicting drug-ATC code associations. Secondly, a detailed examination of the AUROC and AUPR values in Table 8 reveals that the contributions of the six ATC code kernels varied. Notably, the removal of a specific kernel K_{SM}^{i} led to a significant decline in both AUROC and AUPR values. Consequently, according to Table 8, it was determined that kernel K_{SM}^{i} was the most critical. This finding aligns with results from the previous study [27]. In contrast, the models generated by removing any of the other five ATC code kernels exhibited similar AUROC and AUPR values, suggesting that these kernels contributed equally to the overall performance of PDATC-NCPMKL-updated. This highlights the significance of the newly constructed ATC code kernels as being on par with the existing ones.

From these tests, the relative contributions of each ATC code kernel were assessed. Each model utilizing one of the six ATC code kernels was constructed with drug kernels identical to those in PDATC-NCPMKL-updated and subjected to ten-fold cross-validation. This approach facilitates a more direct assessment of the importance of each ATC code kernel. The cross-validation results are summarized in Table 9. It

Table 8 Performance of the models by removing one ATC code kernel

Removed ATC code kernel	Information used for constructing ATC code kernel	Second level		Third level		Fourth level	
		AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
$K_{TP,a}^{i,*}$	Related target proteins of ATC codes	0.9591	0.9640	0.9668	0.9696	0.9694	0.9713
$K_{SE,a}^{i,*}$	Related side effects of ATC codes	0.9593	0.9641	0.9668	0.9695	0.9695	0.9714
$K_{FP,a}^{i,*}$	Related structural features of ATC codes	0.9588	0.9634	0.9660	0.9687	0.9685	0.9709
$K_{F,a}^{i,*\#}$	Known drug-ATC code associations	0.9592	0.9643	0.9648	0.9681	0.9719	0.9724
$K_{SPro}^{i\ \#}$	ATC code tree [24]	0.9598	0.9643	0.9680	0.9706	0.9719	0.9723
$K_{SM}^{i\ \#}$	Numbers or letters in ATC codes [26]	0.9243	0.9451	0.9535	0.9591	0.9679	0.9676

[#] Refer to Additional file 2 for the brief introduction of three ATC code kernels

Table 9 Performance of the models using one ATC code kernel

ATC code kernel	Information used for constructing ATC code kernel	Second level		Third level		Fourth level	
		AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
$K^{i,*}_{TP,a}$	Related target proteins of ATC codes	0.9212	0.9438	0.9518	0.9576	0.9595	0.9626
$K^{i,*}_{SE,a}$	Related side effects of ATC codes	0.9208	0.9436	0.9424	0.9509	0.9553	0.9616
$K^{i,*}_{FP,a}$	Related structural features of ATC codes	0.9191	0.9421	0.9565	0.9618	0.9596	0.9618
$K^{i,*\#}_{F,a}$	Known drug-ATC code associations	0.9258	0.9457	0.9251	0.9402	0.9156	0.9322
$K^{i}_{SPro}{}^{\#}$	ATC code tree [24]	0.7479	0.8133	0.8346	0.8725	0.8592	0.8932
$K^{i}_{SM}{}^{\#}$	Numbers or letters in ATC codes [26]	0.9374	0.9447	0.9484	0.9537	0.9511	0.9573

[#] Refer to Additional file 2 for the brief introduction of three ATC code kernels

can be seen that models using ATC code kernels $K^{i,*}_{TP,a}$, $K^{i,*}_{SE,a}$, $K^{i,*}_{FP,a}$, and K^{i}_{SM} were evidently superior to those using other two ATC kernels ($K^{i,*}_{F,a}$ and K^{i}_{SPro}). The models using ATC code kernels $K^{i,*}_{TP,a}$, $K^{i,*}_{SE,a}$, and $K^{i,*}_{FP,a}$ even yielded better performance than the model using ATC code kernel K^{i}_{SM} at some levels. Thus, it can be concluded that the biological profiles of ATC codes were helpful to predict drug-ATC code associations.

Based on the test results in Tables 8, 9, $K^{i,*}_{TP,a}$, $K^{i,*}_{SE,a}$, $K^{i,*}_{FP,a}$, and K^{i}_{SM} were more important than $K^{i,*}_{F,a}$ and K^{i}_{SPro} for predicting drug-ATC code associations, proving the helpfulness of biological profiles of ATC codes reported in this study.

Performance of the model on different drug-ATC code association groups

This study first inferred the highly related target proteins, structural features, and side effects associated with each ATC code. This information was then integrated into a previously established model to construct a more robust analytical framework. The three factors—target proteins, structural features, and side effects—were found to be closely related to drugs. It is pertinent to explore whether the quantity of related target proteins, structural features, or associated side effects of drugs and ATC codes influences the performance of the model. To investigate this, all drugs were ranked in descending order based on the number of related target proteins. Subsequently, the drugs were equally divided into two groups: the high group and the low group, with the high group containing more related target proteins than the low group. A similar procedure was applied to ATC codes, resulting in two groups (high and low) based on their associated target proteins. The high group for ATC codes also contained a greater number of target proteins compared to the low group. Using the group classifications of drugs and ATC codes, drug-ATC code associations were categorized into four distinct groups: high-high, high-low, low–high, and low-low. For example, the drug-ATC code associations using drugs and ATC codes all in high group constituted the high-high group. For the cross-validation results, we computed the AUROC and AUPR for each of the four drug-ATC code association groups, as detailed in Table 10. The results indicate that the model exhibited nearly equal performance across all four groups at each ATC code level, suggesting limited sensitivity to this factor. This also

Table 10 Performance of PDATC-NCPMKL-updated on different drug-ATC code association groups

Object	Drug group	ATC code group	Second level		Third level		Fourth level	
			AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
Target protein	High	High	0.9629	0.9669	0.9729	0.9743	0.9718	0.9724
	High	Low	0.9631	0.9665	0.9627	0.9657	0.9725	0.9717
	Low	High	0.9636	0.9688	0.9655	0.9692	0.9694	0.9708
	Low	Low	0.9611	0.9665	0.9631	0.9659	0.9733	0.9732
Fingerprint	High	High	0.9633	0.9674	0.9663	0.9685	0.9722	0.9739
	High	Low	0.9670	0.9706	0.9636	0.9680	0.9704	0.9701
	Low	High	0.9574	0.9623	0.9709	0.9723	0.9736	0.9744
	Low	Low	0.9605	0.9660	0.9590	0.9646	0.9699	0.9710
Side effect	High	High	0.9624	0.9677	0.9693	0.9710	0.9744	0.9741
	High	Low	0.9513	0.9534	0.9797	0.9821	0.9759	0.9750
	Low	High	0.9629	0.9668	0.9664	0.9676	0.9679	0.9682
	Low	Low	0.9577	0.9627	0.9689	0.9725	0.9625	0.9638

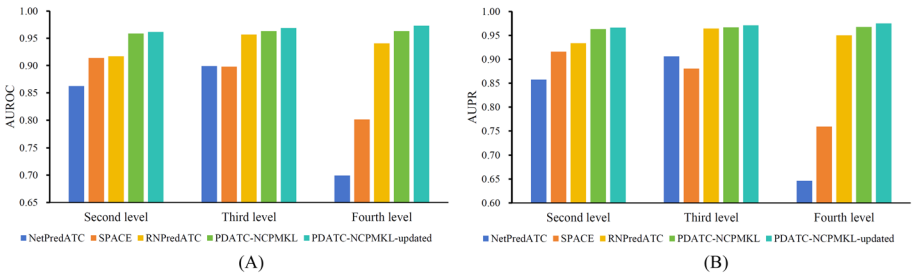

Fig. 6 Bar chart to compare the performance of various models for the prediction of drug-ATC code associations. **A** Bar chart for AUROC; **B** Bar chart for AUPR. The new model PDATC-NCPMKL-updated yields the best performance

confirms the stability of the updated model, PDATC-NCPMKL-updated. Additionally, similar tests were conducted for structural features and side effects, and their corresponding AUROC and AUPR values are also presented in Table 10, leading to the same conclusion.

Comparison with previous models

As indicated in Section "Introduction", several models have been proposed to predict drug-ATC code associations, including NetPredATC [24], SPACE [25], RNPredATC [26], and PDATC-NCPMKL [27]. Figure 6 presents the AUROC and AUPR values for these models. For ease of comparison, the AUROC and AUPR values for PDATC-NCPMKL-updated model are also included in this figure. The analysis reveals that the performance of the earlier models (NetPredATC and SPACE) was relatively poor, with their AUROC values at three levels struggling to exceed 0.9 (Fig. 6A). A similar conclusion can be drawn regarding their AUPR values (Fig. 6B). In contrast, the more recent models (RNPredATC and PDATC-NCPMKL) demonstrated significantly better performance, with all AUROC and AUPR values exceeding 0.9, and reaching as high as 0.95; however, they did not exceed 0.97. The PDATC-NCPMKL-updated model achieved even higher performance, with all AUROC and AUPR values exceeding 0.96, some exceeding

Table 11 Paired student's t-test results of PDATC-NCPMKL-updated and PDATC-NCPMKL

Measurement	Second level	Third level	Fourth level
AUROC	4.227×10^{-4}	1.522×10^{-4}	1.414×10^{-5}
AUPR	3.815×10^{-4}	1.584×10^{-3}	5.817×10^{-6}

Table 12 Performance of the model only with fingerprint-related kernels

ATC code level	AUROC	AUPR
Second level	0.8922	0.9207
Third level	0.9390	0.9502
Fourth level	0.9624	0.9685

0.97. Clearly, the PDATC-NCPMKL-updated model outperformed both RNPredATC and PDATC-NCPMKL, whereas also significantly exceeding NetPredATC and SPACE. Furthermore, a paired Student's t-test was performed comparing PDATC-NCPMKL-updated and PDATC-NCPMKL. The resulting P-values are detailed in Table 11. All P-values were found to be below the 0.05 confidence level, indicating significant differences in performance between PDATC-NCPMKL-updated and PDATC-NCPMKL. Statistically, PDATC-NCPMKL-updated exhibited a significant advantage over PDATC-NCPMKL.

The updated model serves as an enhancement of the previous PDATC-NCPMKL model. By incorporating the biological profiles of ATC codes, the ATC code kernels used in PDATC-NCPMKL-updated provide much more informative data compared to those utilized in PDATC-NCPMKL. This increased informativeness is the primary reason for the superiority of PDATC-NCPMKL-updated over PDATC-NCPMKL. In contrast, the three earlier models employed significantly less information regarding drugs and ATC codes, which hindered their ability to adequately evaluate or represent these entities, resulting in lower performance compared to PDATC-NCPMKL-updated.

Performance of the model only with fingerprint-related kernels

This study enhances an existing drug-ATC code association prediction model by incorporating biological profiles of ATC codes. As a result, the updated PDATC-NCPMKL model requires several specific properties of both drugs and ATC codes, which may restrict its applicability. If certain properties of either drugs or ATC codes are unavailable, the PDATC-NCPMKL-updated model may not provide reliable prediction results. Drug fingerprints, which represent the essential structures of drugs, are commonly used for this purpose. Some earlier methods were developed using only the structural information of drugs [13, 22, 24, 26]. It is noteworthy to evaluate our model using solely fingerprint-related kernels. In the PDATC-NCPMKL-updated model, both the drug kernel $K_{FP,d}^{i,*}$ and the ATC code $K_{FP,a}^{i,*}$ were derived from the drugs' ECFP fingerprint. The model, utilizing these two kernels along with the association adjacency matrix F^i, was constructed and assessed through ten-fold cross-validation. The test results are presented in Table 12. The findings indicate that the model's performance exhibited an upward trend with increasing ATC code levels. When compared to the performance of

the PDATC-NCPMKL-updated model (Section "Performance of PDATC-NCPMKL-updated" and Fig. 5), the differences in AUROC and AUPR at the second level ranged from approximately 4% to 7%. This gap decreased to about 2% to 3% at the third level and further narrowed to around 1% at the fourth level. This suggests that whereas the model using only fingerprint-related kernels was less effective than the PDATC-NCPMKL-updated model, it nonetheless demonstrates a broader application scope.

To validate this performance, two prior models (NetPredATC [24] and RNPredATC [26])—which relied solely on drug structural information and were designed for predicting drug-ATC code associations—were chosen for comparison. A bar chart illustrating the AUROC and AUPR of the three models is shown in Fig. 7. The results indicate that the model employing only fingerprint-related kernels outperformed NetPredATC. In comparison to RNPredATC, the model with fingerprint-related kernels exhibited slightly lower performance at the second level, approximately equal performance at the third level, and superior performance at the fourth level. Thus, the model utilizing only fingerprint-related kernels proves to be competitive with earlier models, indicating its potential as a beneficial tool for predicting drug-ATC code associations.

Limitations and applications of PDATC-NCPMKL-updated

The PDATC-NCPMKL-updated model incorporates several properties of drugs and ATC codes. Although this approach enhances its performance, it also introduces certain limitations. The requirement for multiple properties during model execution may restrict its applicability. For example, if the target proteins or side effects of a drug are unavailable, the PDATC-NCPMKL-updated model may not produce reliable predictions for associations involving that drug. In future, we aim to address this limitation in our ongoing work.

PDATC-NCPMKL-updated demonstrates several valuable applications. Firstly, its cross-validation results indicate that false positive predictions may reveal latent drug-ATC code associations, potentially leading to the assignment of new ATC codes to existing drugs and identifying new diseases that these drugs may treat. Secondly, the trained PDATC-NCPMKL-updated can be employed to evaluate unlabeled drug-ATC code pairs. A positive test result in this context also suggests a novel ATC code for a drug. Overall, PDATC-NCPMKL-updated offers valuable insights into the discovery of new therapeutic and pharmacological effects of drugs.

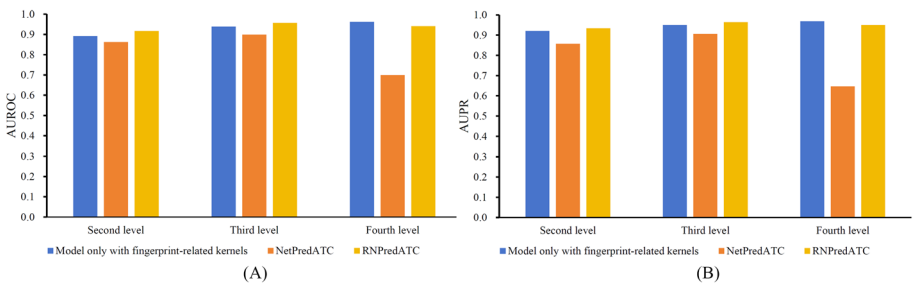

Fig. 7 Bar chart to compare the performance of various models only using drug structure information for the prediction of drug-ATC code associations. **A** Bar chart for AUROC; **B** Bar chart for AUPR. The model with fingerprint-related kernels, derived from PDATC-NCPMKL-updated, provides competitive performance

Conclusions

The drug ATC classification system is a critical area of research in drug repositioning. This study presents an updated model of PDATC-NCPMKL that incorporates newly inferred biological profiles of ATC codes. The updated model demonstrated superior performance compared to the original PDATC-NCPMKL, highlighting the positive impact of the inferred biological profiles of ATC codes on predicting drug-ATC code associations. The enhanced efficacy of the new model is further validated by its performance against other existing models. Additionally, the importance of the biological profiles of ATC codes was confirmed through ablation tests on the ATC code kernels. It is optimistic that this new model will serve as a valuable tool for exploring the ATC classification system, and that the inferred biological profiles of ATC codes can contribute to addressing other drug-related challenges. The codes and data associated with this study are available at https://github.com/Lywhere/PDATC-NCPMKL-updated.

Abbreviations

ATC	Anatomical therapeutic chemical
R&D	Research and development
WHO	World Health Organization
RWR	Random walk with restart
SMILES	Simplified molecular input line entry system
WKNKN	Weighted K nearest known neighbors

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06102-7.

Additional file 1.

Additional file 2.

Additional file 3.

Additional file 4.

Additional file 5.

Acknowledgements
Not applicable.

Author contributions
L.C. designed the research; L.C., Y.L. and J.X. conducted the experiments; Y.L. and B.Z. analyzed the results. All authors have read and approved the manuscript.

Funding
Not applicable.

Availability of data and materials
The source codes and data are available at https://github.com/Lywhere/PDATC-NCPMKL-updated.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 9 August 2024 Accepted: 4 March 2025
Published online: 21 March 2025

References

1. Luo H, Li M, Yang M, Wu FX, Li Y, Wang J. Biomedical data and computational models for drug repositioning: a comprehensive review. Brief Bioinform. 2021;22(2):1604–19.
2. Dowden H, Munro J. Trends in clinical success rates and therapeutic focus. Nat Rev Drug Discov. 2019;18(7):495–6.
3. Jarada TN, Rokne JG, Alhajj R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. J Cheminform. 2020;12(1):46.
4. Govindaraj RG, Naderi M, Singha M, Lemoine J, Brylinski M. Large-scale computational drug repositioning to find treatments for rare diseases. NPJ Syst Biol Appl. 2018;4:13.
5. Nahler G, Nahler G. Anatomical therapeutic chemical classification system (ATC). Dict Pharm Med 2009:8–8.
6. Chen L, Zeng W-M, Cai Y-D, Feng K-Y, Chou K-C. Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. PLoS ONE. 2012;7(4):e35254.
7. Chen L, Lu J, Zhang N, Huang T, Cai Y-D. A hybrid method for prediction and repositioning of drug anatomical therapeutic chemical classes. Mol BioSyst. 2014;10(4):868–77.
8. Cheng X, Zhao S-G, Xiao X, Chou K-C. iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. Bioinformatics. 2016;33(3):341–6.
9. Cheng X, Zhao SG, Xiao X, Chou KC. iATC-mHyb: a hybrid multi-label classifier for predicting the classification of anatomical therapeutic chemicals. Oncotarget. 2017;8(35):58494–503.
10. Nanni L, Brahnam S. Multi-label classifier based on histogram of gradients for predicting the anatomical therapeutic chemical class/classes of a given compound. Bioinformatics. 2017;33(18):2837–41.
11. Wang X, Wang Y, Xu Z, Xiong Y, Wei DQ. ATC-NLSP: prediction of the classes of anatomical therapeutic chemicals using a network-based label space partition method. Front Pharmacol. 2019;10:971.
12. Zhou J-P, Chen L, Guo Z-H. iATC-NRAKEL: An efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs. Bioinformatics. 2020;36(5):1391–6.
13. Zhou J-P, Chen L, Wang T, Liu M. iATC-FRAKEL: a simple multi-label web-server for recognizing anatomical therapeutic chemical classes of drugs with their fingerprints only. Bioinformatics. 2020;36(11):3568–9.
14. Lu Z, Chou K-C. iATC_Deep-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals by deep learning. Adv Biosci Biotechnol. 2020;11(5):153–9.
15. Zhao H, Li Y, Wang J. A convolutional neural network and graph convolutional network-based method for predicting the classification of anatomical therapeutic chemicals. Bioinformatics. 2021;37(18):2841–7.
16. Tang S, Chen L. iATC-NFMLP: identifying classes of anatomical therapeutic chemicals based on drug networks, fingerprints and multilayer perceptron. Curr Bioinform. 2022;17(9):814–24.
17. Lumini A, Nanni L. Convolutional neural networks for ATC classification. Curr Pharm Des. 2018;24(34):4007–12.
18. Nanni L, Brahnam S, Lumini A. Ensemble of deep learning approaches for ATC classification. In: Smart intelligent computing and applications. Springer; 2020, pp. 117–125.
19. Yan C, Suo Z, Wang J, Zhang G, Luo H. DACPGTN: drug ATC code prediction method based on graph transformer network for drug discovery. Front Pharmacol. 2022;13:907676.
20. Nanni L, Lumini A, Brahnam S. Neural networks for anatomical therapeutic chemical (ATC) classification. Appl Comput Inf. 2022, ahead-of-print(ahead-of-print).
21. Wang X, Liu M, Zhang Y, He S, Qin C, Li Y, Lu T. Deep fusion learning facilitates anatomical therapeutic chemical recognition in drug repurposing and discovery. Brief Bioinform. 2021;22(6):bbab289.
22. Cao Y, Yang Z-Q, Zhang X-L, Fan W, Wang Y, Shen J, Wei D-Q, Li Q, Wei X-Y. Identifying the kind behind SMILES—anatomical therapeutic chemical classification using structure-only representations. Brief Bioinf. 2022;23:bbac346.
23. Chen FS, Jiang ZR. Prediction of drug's anatomical therapeutic chemical (ATC) code by integrating drug-domain network. J Biomed Inform. 2015;58:80–8.
24. Wang YC, Chen SL, Deng NY, Wang Y. Network predicting drug's anatomical therapeutic chemical code. Bioinformatics. 2013;29(10):1317–24.
25. Liu Z, Guo F, Gu J, Wang Y, Li Y, Wang D, Lu L, Li D, He F. Similarity-based prediction for anatomical therapeutic chemical classification of drugs by integrating multiple data sources. Bioinformatics. 2015;31(11):1788–95.
26. Zhao H, Duan G, Ni P, Yan C, Li Y, Wang J. 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):2712–23.
27. Chen L, Xu J, Zhou Y. 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.
28. Peng Y, Wang M, Xu Y, Wu Z, Wang J, Zhang C, Liu G, Li W, Li J, Tang Y. Drug repositioning by prediction of drug's anatomical therapeutic chemical code via network-based inference approaches. Brief Bioinform. 2020;22(2):2058–72.
29. Chen L, Liu T, Zhao X. Inferring anatomical therapeutic chemical (ATC) class of drugs using shortest path and random walk with restart algorithms. BBA Mol Basis Dis. 2018;1864(6, Part B):2228–40.
30. Liang HY, Hu B, Chen L, Wang SQ. Aorigele: recognizing novel chemicals/drugs for anatomical therapeutic chemical classes with a heat diffusion algorithm. Bba-Mol Basis Dis. 2020;1866(11):165910.
31. Tong H, Faloutsos C, Pan J. Fast random walk with restart and its applications. In: Sixth international conference on data mining (ICDM'06): 18–22 Dec. 2006, vol. 2006, pp. 613–622.
32. Kohler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet. 2008;82(4):949–58.
33. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074–82.
34. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34(suppl 1):D668–72.
35. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28:31–6.
36. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–54.
37. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016;44(D1):D1075-1079.

38. Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res. 2007;36(suppl_1):D684–8.
39. Kuhn M, Szklarczyk D, Pletscher-Frankild S, Blicher TH, von Mering C, Jensen LJ, Bork P. STITCH 4: integration of protein-chemical interactions with user data. Nucleic Acids Res. 2013;42(D1):D401–7.
40. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva Nadezhda T, Pyysalo S, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2022;51(D1):D638–46.
41. Huang F, Guo W, Chen L, Feng K, Huang T, Cai Y-D. Identifying autophagy-associated proteins and chemicals with a random walk-based method within heterogeneous interaction network. Front Biosci-Landmark. 2024;29(1):21.
42. Li L, Wang Y, An L, Kong X, Huang T. A network-based method using a random walk with restart algorithm and screening tests to identify novel genes associated with Menière's disease. PLoS ONE. 2017;12(8):e0182592.
43. Zhang Y, Dai L, Liu Y, Zhang Y, Wang S. Identifying novel fruit-related genes in *Arabidopsis thaliana* based on the random walk with restart algorithm. PLoS ONE. 2017;12(5):e0177017.
44. Guo W, Shang D-M, Cao J-H, Feng K, He Y-C, Jiang Y, Wang S, Gao Y-F. Identifying and analyzing novel epilepsy-related genes using random walk with restart algorithm. Biomed Res Int. 2017;2017:13.
45. Ezzat A, Zhao P, Wu M, Li XL, Kwoh CK. Drug-target interaction prediction with graph regularized matrix factorization. IEEE/ACM Trans Comput Biol Bioinform. 2017;14(3):646–56.
46. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International joint conference on artificial intelligence, 1995, Lawrence Erlbaum Associates Ltd, pp. 1137–1145.
47. Powers D. Evaluation: from precision, recall and f-measure to roc., informedness, markedness & correlation. J Mach Learn Technol. 2011;2(1):37–63.
48. Chen L, Li J. PDTDAHN: predicting drug-target-disease associations using a heterogeneous network. Curr Bioinf. 2025.
49. Chen L, Gu J, Zhou B. PMiSLocMF: predicting miRNA subcellular localizations by incorporating multi-source features of miRNAs. Brief Bioinf. 2024;25(5):bbae86.
50. Chen L, Chen Y. RMTLysPTM: recognizing multiple types of lysine PTM sites by deep analysis on sequences. Brief Bioinf. 2024;25(1):bbad450.

Publisher's Note